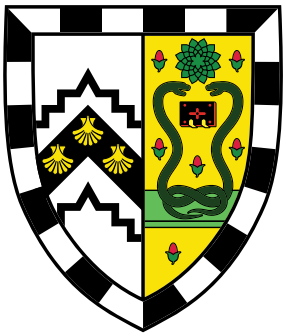
Computationally Efficient Methods for High-Dimensional Statistical Problems



Benjamin George Stokell

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville and Caius College July 2021

For my parents.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. It does not exceed the prescribed word limit for the Mathematics Degree Committee.

Chapter 2 is joint with with Rajen Shah (University of Cambridge) and Ryan Tibshirani (Carnegie Mellon University), and has been published in the *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* as Stokell et al. (2021b). Chapters 3 and 4 are joint work with Rajen Shah (University of Cambridge). Chapter 3 constitutes work I intend to submit for publication in the near future.

Benjamin George Stokell
July 2021

Acknowledgements

I would first like to thank Rajen Shah, my supervisor, for all of his guidance and insight throughout my PhD in which I have learnt a great deal. Research is inherently an endeavour full of uncertainty and I am enormously grateful for the generosity of his time, ideas and mentorship. I would also like to thank my collaborator, Ryan Tibshirani, with whom I enjoyed working on the second chapter. I would like to express my gratitude to Richard Samworth, for inviting me to join his group meetings during my first year, and for his warm welcome to the Statistical Laboratory in general.

I have enjoyed valuable discussions with colleagues in Rajen's group, including Harvey Klyne, Anton Lundborg, Florian Pein and Yuhao Wang. My extended office mates, Derek Driggs, Tom Edinburgh, Oliver Feng, Tamara Großmann, Torben Sell and Maggie Steward, have been supportive and insightful and a pleasure to work alongside. I'll treasure many memories of our time in the basement of Pavilion F. I would also like to thank my friends in the department for their company and conviviality, including many of those listed already and also Benjamin Barrett, James Bell, Tom Berrett, Nigel Burke, Nicolas Dupré, Jo Evans, James Kilbane and James Munro.

I much enjoyed the two years I spent serving on the Gonville & Caius MCR committee and would like to thank those I worked with more closely including Anthony Bridgen, Jessica Corry and Lindsay Pike. I am particularly grateful for the support of friends including Elizabeth Harper and Charles McCombe during the more challenging moments of working towards a PhD. Thanks also to those who unwittingly became my office mates since the start of the COVID-19 pandemic, Edward Beard, Amelia Drew, Joep ten Wolde and Peter van Hintum. I am thankful to the CCIMI for providing funding without which the work contained in this thesis would not have been possible, and for the friendly and relaxed environment that made it so enjoyable.

Thank you to my parents as I would not be in this position but for their love and encouragement over many years. Finally, I am deeply grateful to Navoditte Das, whose unwavering kindness, support and patience has greatly improved the state of this thesis and my life.

Abstract

With the ever-increasing amount of computational power available, so broadens the horizon of statistical problems that can be tackled. However, many practitioners have only an ordinary personal computer on which to do their work. The need for computationally efficient methodology is as pressing as ever, and there remain some questions as-yet without a confident answer for a practitioner working with tight computational constraints. This thesis develops methods for three such problems. The first, introductory, chapter provides an overview of the area and an accessible preamble to the problems these methods address.

In the second chapter we address the problem of modelling a high-dimensional linear regression with categorical predictor variables. The natural sparsity assumption in this setting is on the number of unique values the coefficients within each categorical variable can take. With this assumption, we introduce a new form of penalty function for tackling this problem. While the number of combinations of levels can grow extremely fast in the number of levels, the unique structure of the method enables fast optimisation for this problem. A novel and intricate dynamic programming algorithm computes the exact global optimum over each variable, and is embedded within a block coordinate descent algorithm. This allows fitting of such models quickly on a laptop computer in a memory efficient manner. The scaling requirements sufficient for this method to recover the correct groups cannot be relaxed for any estimator; this strong performance is validated by a range of experiments using both simulated and real data.

In the third chapter we explore the possibility that a practitioner has some a priori belief to which variables are most likely to be important, which will be in the form of a permutation of the columns. Our approach takes this ordering and efficiently computes a grid of solution paths by sequentially removing groups of variables without unnecessary recomputation of coefficients. Typical examples of such orderings include the column norms in the (unscaled) design matrix, or the recentness of observations in time series data. This procedure, combined with selecting the size of support set by validation on a test set, has similar performance to that of fitting the oracular submodel.

The fourth chapter concerns the efficient estimation of conditional independence graphs in Gaussian graphical models. Neighbourhood selection (Meinshausen and Bühlmann, 2006) is practical, popular, and enjoys good performance, but in large-scale settings it can still have

computational demands exceeding the resources available to many practitioners. Screening approaches promise large improvements in speed with only a small price to pay in terms of resulting estimation performance. Although it is well-known that nodes adjacent in the conditional independence graph may be uncorrelated, a minimum absolute correlation between adjacent nodes is often tacitly or explicitly assumed in order for screening procedures to be effective. We make use of recent work in covariance estimation and high-dimensional screening of variables to develop a fast, two-stage, screening procedure specifically for use within neighbourhood selection and avoiding this restrictive assumption. Provided that a weaker version of a minimum edge strength requirement holds over most of the graph, the performance of the post-screening nodewise regressions is not compromised, while being substantially faster than the full procedure. This method is robust to the presence of latent confounders, as well as other scenarios that typically impede the screening of variables. Experiments show that our approach strikes a favourable balance between edge detection and computational efficiency.

Contents

Chapter 1

Introduction

For roughly a century, likelihood-based estimation and inference (Fisher, 1922) has been central to much of statistical practice and theory. In a regression setting, we suppose that the observations $(Y_i, X_i)_{i=1}^n$ are drawn independently such that $Y_i|X_i \sim f(\cdot; X_i, \theta)$, where $\theta \in \Theta$ is the unknown parameter in some space Θ, and X_i is independent of θ. The likelihood is then

$$l(\theta; (X_i, Y_i)_{i=1}^n) = \prod_{i=1}^n f(Y_i; X_i, \theta),$$

as a function of θ. The maximum likelihood estimate $\hat{\theta}$ is the maximiser of $l(\theta; (X_i, Y_i)_{i=1}^n)$ over $\theta \in \Theta$. Within this very general approach, it is often appropriate and practical to work in the generalised linear model framework. This is a class of parametric models where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $Y_i|\boldsymbol{x}_i \sim f(\cdot\,; \mu_i, \sigma^2)$, where it is assumed $g(\mu_i) = \boldsymbol{x}_i^T\beta$ for some specified function g, and with parameters $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_{>0}$.

Let us now consider the example of the normal linear model, where $Y_i \sim \mathcal{N}(\boldsymbol{x}_i^T\beta, \sigma^2)$ independently. The estimate $\hat{\beta}$ of β is selected by minimising the negative log-likelihood, or equivalently

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2, \tag{1.1}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times p}$ and $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T \in \mathbb{R}^n$, which in this case coincides with the least squares estimate. Provided $\boldsymbol{X}$ is of full column rank, this solution has closed form $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, with distribution $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. This is the basis for inference on β, which encompasses testing the significance of individual coefficients or entire models.

In the setting where $p > n$, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ will not be of full rank and thus cannot be inverted. Minimising the squared loss in (1.1) results in an underdetermined linear problem

with a subspace of solutions, so the least squares objective does not have a unique minimiser. This problem is an example of the methodological and theoretical complications that can arise in the *high-dimensional setting*, where p is not of smaller order than n. This thesis makes contributions to methodology for a selection of practical and important problems related to regression in the high-dimensional setting.

1.1 Background

Many traditional methods for linear modelling in the high-dimensional setting are based around selecting a small subset of variables, thus returning to the low-dimensional setting in which estimation and prediction can be performed by classical procedures such as ordinary least squares. After fitting a (potentially quite large) number of models on low-dimensional subsets of variables, one will be selected; Miller (1990) provides an overview of subset selection approaches.

Having computed estimates $(\hat{\beta}(S), \hat{\sigma}^2(S))$ for each model S (with dimension d) and denoting $\hat{\ell} = \log(l(\hat{\beta}(S), \hat{\sigma}^2(S); (\boldsymbol{x}_i, Y_i)_{i=1}^n))$, this can be done by maximising an information criterion such as Akaike Information Criterion (AIC), $2(d - \hat{\ell})$ (Akaike, 1974), or Bayesian Information Criterion (BIC), $2(\frac{1}{2}d\log(n) - \hat{\ell})$ (Schwarz, 1978) (see e.g. Chen and Chen (2008); Wang et al. (2009) and Kim et al. (2012) for some more recent perspectives). These manage the tradeoff between the flexibility associated with fitting a larger model with the increased tendency of such a model to overfit to the data. Such criteria can often be characterised as adding a penalty term to the negative log-likelihood (1.1) to discourage the inclusion of too many variables, but as these can be too difficult to optimise directly, their use remains primarily for model comparison. The strategy of forward regression (see e.g. Wang (2009)) which entails including variables sequentially by a greedy selection criterion, provides a sequence of at most $\min\{p, n\}$ models for comparison. This is much less computationally demanding than exhaustively fitting all low-dimensional submodels.

Another approach for estimation of β in the high-dimensional linear model is to directly remedy the non-invertibility of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ by adding to it a multiple of the identity matrix, yielding the ridge regression estimate (Hoerl and Kennard, 1970; Tikhonov, 1963):

$$\hat{\beta}^{\text{ridge},\lambda} = (\boldsymbol{X}^T\boldsymbol{X} + n\lambda I_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

where $\lambda > 0$ is a tuning parameter. This explicitly bounds the minimum eigenvalue of $(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p)$ away from zero, guaranteeing the existence of the estimator. Note that this is the solution of a penalised version of the ordinary least squares objective:

$$\hat{\beta}^{\text{ridge},\lambda} = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}, \tag{1.2}$$

where the penalty term $\|\beta\|_2^2$ has the effect of shrinking the solution vector, in particular large coefficients, towards 0. This reduces the variance of the estimate (following similar intuition to the celebrated James–Stein estimator (James and Stein, 1961; Samworth, 2005)), improving its mean squared prediction error even in low dimension. This is an example of *bias-variance tradeoff*, an overarching theme of efforts for estimation in high-dimensional settings.

While ridge regression enables the convenient fitting of high-dimensional models, it is difficult to interpret its individual coefficients. Identifiability of the true model in the high-dimensional setting requires further structural assumptions on β, the most popular of which is *sparsity*. This states that of all the p coefficients, only $s \ll p$ are non-zero, typically meaning that the underlying model is low-dimensional and the remaining variables whose coefficients are zero are redundant. In this context, ridge regression is less attractive (Frank and Friedman, 1993) than alternatives which by some means select a low-dimensional model, compared to only shrinking the ℓ_2 norm of the coefficients.

By the mid-1990s a range of new approaches had been proposed for simultaneous variable selection and model fitting, including those based on Gibbs sampling (George and McCulloch, 1993) and also the 'non-negative garrotte' estimator (Breiman, 1995; Zou, 2006). The non-negative garotte provides a continuous optimisation objective for both variable selection and coefficient shrinkage, but is in essence a scaled version of the ordinary least squares estimate (with some coefficients scaled all the way to zero) and therefore not applicable in the high-dimensional setting. The proposal which gained by far the most traction is the Lasso (least absolute shrinkage and selection operator) (Tibshirani, 1996):

$$\hat{\beta}^{\text{lasso},\lambda} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}. \tag{1.3}$$

Its applicability in the high-dimensional setting, not to mention its arguably more straightforward construction than e.g. the non-negative garotte, quickly established it as the state of the art. Furthermore, its objective is strictly convex in β (provided $\lambda > 0$) and can therefore be efficiently solved, for example by cyclical coordinate descent. Alternatives include the Dantzig selector (Candes and Tao, 2007)

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1, \quad \text{subject to: } \|\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\beta)\|_\infty \leq \lambda,$$

which uses a different formulation to the penalised least squares optimisation objective (1.4) below, but ultimately has similar statistical properties to the Lasso (Bickel et al., 2009).

Similarly to ridge regression (1.2), the Lasso (1.3) solves a penalised least squares optimisation problem of the form

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \mathcal{P}(\beta) \right\}, \tag{1.4}$$

where $\mathcal{P}(\beta)$ is a penalty term. In each of these objectives, this plays the role of a *regularisation* term. Regularisation is useful not just in making high-dimensional problems solvable in the first place, but also in playing the crucial role of managing the flexibility of the model and preventing overfitting. For this reason it is a valuable and general tool in the statistician's toolbox for controlling the bias-variance tradeoff. For example, both AIC and BIC are equivalent to a penalty function

$$\mathcal{P}(\beta) \propto \|\beta\|_0 = |\{j \colon \beta_j \neq 0\}|.$$

Such ℓ_0-penalised estimators are of interest in their own right but, as mentioned previously, can be difficult to optimise directly, though work has been done in developing strategies for this problem (see Huang et al. (2018) and references therein). For more computational or machine learning-focused perspectives on regularisation in general, see Chapters 7 of Parikh and Boyd (2014) or Goodfellow et al. (2016) respectively.

The Lasso is able to set entries exactly to zero because it has the *selection property* for each variable j, namely that the one-sided partial derivative with respect to individual coefficients at zero is strictly positive. The KKT conditions (Karush, 1939; Kuhn and Tucker, 1951) on the subgradient of the objective (1.3) imply that all variables j such that $|\frac{1}{n}\boldsymbol{X}_j^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})| < \lambda$ must be deselected. This allows simultaneous variable selection and model fitting, but sometimes a separate dimension reduction step is warranted. In ultrahigh dimensions, e.g. where $\log(p)/n \not\to 0$, even methods such as the Lasso will not perform well, but in practice the primary motivation for rapidly reducing the number of variables will be computational. A popular approach for this is screening variables based on their marginal correlation with the response (Fan and Lv, 2008), an approach which can be extended to a wide range of model settings (Fan et al., 2009, 2011; Gorst-Rasmussen and Scheike, 2013).

The strong empirical performance of the Lasso has seen its use flourish in practice and theoretical guarantees have been developed to corroborate this. There is a vast literature on this topic, with textbooks such as Bühlmann and van de Geer (2011); Hastie et al. (2016) and Wainwright (2019) providing clear overviews of the area. Variable selection remains more challenging than prediction and estimation, with the assumptions required for selection consistency (Meinshausen and Yu, 2009; Zhao and Yu, 2006) being stronger than those for prediction and estimation (van de Geer and Bühlmann, 2009). The performance of the Lasso for variable selection can be improved by use of subsampling strategies (Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013).

The task of performing inference on the coefficients of Lasso solutions is a formidable technical challenge, owing to the intractability of the distribution of the Lasso estimate. Following an ordinary least squares approach on the low-dimensional submodel consisting of the set of variables selected by an initial Lasso estimate (Efron et al., 2004; Meinshausen, 2007) is not suitable since the selection step constitutes 'data snooping', though Belloni et al. (2014) presents a strategy for doing so in a valid way. The post-selection inference literature provides a range of methods for correcting for the influence of the selection process, both for specific models such as the Lasso (Lee et al., 2016; Lockhart et al., 2014; Tibshirani et al., 2016) and after a general model selection procedure (Berk et al., 2013). A breakthrough came in the work of Javanmard and Montanari (2014); Zhang and Zhang (2014) and van de Geer et al. (2014), which provides an approximate distribution for the entire p-dimensional Lasso estimate. They proceed by removing an estimate of the bias from coefficients to obtain (under certain conditions, and in a setting where the rows of X are random) distributions approximately centred around each of the true coefficients,

$$\hat{\beta}_j + \frac{1}{n}\hat{\Omega}_j \boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \beta_j + \frac{1}{n}\hat{\Omega}_j \boldsymbol{X}^T\varepsilon + o_P(n^{-1/2}). \tag{1.5}$$

Here, $\hat{\Omega}_j$ is an estimate of the j^{th} row of the inverse of Σ, the covariance matrix for rows $\boldsymbol{x}_i$ of the design matrix $\boldsymbol{X}$. We will discuss different strategies for constructing these estimates in Section 1.1.2.

In the wake of the popularity of the Lasso, a vast literature of methodology has grown, heavily exploiting the malleability that the choice of penalty term provides over the type of shrinkage encouraged in the solution. In the presence of highly correlated covariates, it is observed that the Lasso solution will typically select at most one, in an unstable manner. This can be addressed by instead using the 'elastic net' penalty (Zou and Hastie, 2005; Zou and Zhang, 2009), a convex combination of both Lasso and ridge penalties,

$$\mathcal{P}(\beta) = \lambda\left(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2\right),$$

where $\alpha \in [0, 1]$. Alternatively, 'SLOPE' (Bogdan et al., 2015) uses non-uniform weights for each of the absolute value penalty terms.

Another common complaint is that too much shrinkage is applied to the non-zero coefficients in Lasso solutions, leading to their estimates being excessively biased towards zero. This behaviour can be mitigated by refitting the Lasso with a smaller penalty parameter just over those variables selected by the first Lasso fit (Meinshausen, 2007), or doing so with a reweighted version, reducing the penalty on those variables with larger initial coefficients

(Zou, 2006),

$$\mathcal{P}(\beta; \hat{\beta}^{\text{init}}) = \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_j^{\text{init}}|}.$$

Note that this is taken to place an infinite penalty on those variables not selected by the initial Lasso solution $\hat{\beta}^{\text{init}}$. In practice, this simply means that such variables are excluded from the model before one proceeds to compute the solution. The tendency to overly shrink non-zero coefficients is also addressed by a family of penalty functions which replace the absolute value terms $|\beta_j|$ in the penalty with some concave function ρ thereof (Zhang and Zhang, 2012),

$$\mathcal{P}(\beta) = \lambda \sum_{j=1}^{p} \rho(|\beta_j|). \tag{1.6}$$

For example, the bridge regression procedure (Frank and Friedman, 1993; Fu, 1998) specifies $\rho(|\beta_j|) = |\beta_j|^\gamma$ for $\gamma \geq 0$, of which both ridge ($\gamma = 2$) and Lasso ($\gamma = 1$) penalties are special cases, is concave for choices $\gamma < 1$. Piecewise quadratic constructions such as 'SCAD' (Fan and Li, 2001) and 'MCP' (Zhang, 2010) can be favourable due to the zero shrinkage they impose on coefficients whose size exceeds a certain threshold.

This use of nonconvex penalties can have favourable variable selection properties (Loh and Wainwright, 2017), but can bring significant computational challenges that have been addressed in some more recent and often highly technical work (Fan et al., 2018; Loh and Wainwright, 2015; Wang et al., 2014; Zhao et al., 2018). These typically proceed either by arguing that all local optima share favourable properties, or instead that 'good' local optima have a sufficiently large basin of attraction that the solution of a similar, but convex, problem falls inside it.

1.1.1 Structured problems

In many problems, covariates will not be structurally indistinguishable, by which we mean that they are endowed with some structural information informing the nature of the variable selection that should be applied. Many of the concepts behind the various methods for vanilla regression problems can find new applications, here motivated by the additional assumptions that may be made on the coefficient vector. To make this concrete, we will consider some notable examples.

Covariates may naturally arise in groups $\mathcal{G}_1, \ldots, \mathcal{G}_J$, and in such a setting it may be appropriate to set entire groups of variables to zero simultaneously. An example of this is a categorical variable, whose levels are encoded by a group of binary dummy variables, so its deselection is equivalent to setting all of these dummy variables to zero. The 'group Lasso'

approach (Yuan and Lin, 2006),

$$\mathcal{P}(\beta) = \lambda \sum_{g=1}^{J} \|\beta_{\mathcal{G}_g}\|_2,$$

provides a way to do this. The effect of the terms in the penalty is analogous to that of the absolute value terms in the Lasso, but shrinking all variables inside each group towards zero. This can be combined with the usual Lasso penalty to also encourage coefficientwise sparsity (Simon et al., 2013) and other modifications exist to address more general structured sparsity settings, such as potentially overlapping group structures (Bach et al., 2011; Jacob et al., 2009; Jenatton et al., 2011).

Another example of a type of structural information over the variables is adjacency. This could be spatial (regions sharing a border), temporal (variables being adjacent in time), or any other notion that is applicable to the setting. The need to somehow reduce the dimensionality of the model remains, but naïvely sending a lot of the coefficients to zero leaves an inflexible model that is not easily interpretable. Instead the 'fused Lasso' (Tibshirani et al., 2005; Tibshirani and Taylor, 2011) allows a graph to be placed over all of the variables, with a penalty

$$\mathcal{P}(\beta) = \lambda \sum_{(j,k)\in E} |\beta_j - \beta_k|$$

including terms to encourage the difference between coefficients for adjacent nodes (variables) to become zero, thereby fusing them together.

Returning to the setting of categorical covariates, in large-scale settings the model selected by the group Lasso may remain high-dimensional, with only a shrinkage (rather than selection) penalty imposed within the subvector corresponding to levels of a variable that has been selected. Further reduction in model dimension necessitates some levels within the selected categorical variables to be collapsed together, i.e. fused such that they share a coefficient. The sparse group Lasso is only able to shrink the dummy coefficients towards zero, precluding a symmetric treatment of the levels.

The fused Lasso provides a natural way to do this in the case of ordinal categorical covariates, as it is clear which levels are adjacent. In fact, there is no need to include an additional group Lasso term in the penalty: the fused Lasso is capable of fusing all levels within a categorical variable together, which has the effect of deselecting the variable because it is absorbed into the intercept. However, if the estimated coefficients are ordered consistently with the given ordering of the ordinal variable, then the penalty terms can telescope and only penalise the range of the coefficients (Oelker et al., 2015).

In the case of nominal categorical covariates, a natural procedure to consider is another fused Lasso penalty, this time with respect to the complete graph, i.e. every pair of levels

within a variable are connected by an edge. This is the 'CAS-ANOVA' penalty introduced explicitly in Bondell and Reich (2009),

$$\mathcal{P}(\beta) = \lambda \sum_{j=1}^{p} \sum_{k_1=1}^{K_j-1} \sum_{k_2=k_1+1}^{K_j} w_{j,k_1k_2} |\beta_{jk_1} - \beta_{jk_2}| \tag{1.7}$$

(where variable j has categories $1, \ldots, K_j$), along with a scheme for weighting penalty terms in accordance with the frequencies of different levels in the training data. The combination of these approaches for data with both nominal and ordinal variables is explored in Gertheiss and Tutz (2010). Related penalties include replacing the absolute value terms in (1.7) with some function thereof. For example, Ma and Huang (2017) consider concave functions for the same reasons as in the vanilla setting (1.6).

This 'all-pairs' approach is undesirable for a number of reasons. Firstly, it increases the number of terms in the penalty function dramatically such that optimisation is more involved, often requiring a large number of slack variables in order to impose the penalty. Also, the amount of shrinkage between coefficients depends on the estimated sparsity pattern in such a way that collapsed levels of very uneven size are encouraged. Addressing this problem is the subject of Chapter 2.

1.1.2 Precision matrix estimation

Suppose now that we are in a setting where all of the covariates are continuous and structurally indistinguishable, and we assume that the rows $\boldsymbol{x}_i$ are drawn independently from a distribution with covariance Σ. Here if we have a response vector $\boldsymbol{Y}$ it can be discarded and for each variable j, we can consider the j^{th} column $\boldsymbol{X}_j$ to be the response variable upon which we carry out a 'nodewise regression'

$$\hat{\beta}^j \in \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}. \tag{1.8}$$

Examples of this setting include the results of gene sequencing with unlabelled donors, as we explore in Chapter 4. This can be shown (Lauritzen, 1996) to estimate the columns of the 'precision matrix' $\Omega = \Sigma^{-1}$ of the distribution from which rows of the design matrix are drawn, up to a scale factor. This follows from the population minimiser of (1.8) satisfying

$$\underset{\beta \in \mathbb{R}^{p-1}}{\arg\min}\, \mathbb{E}\left[\frac{1}{2} \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\beta\|_2^2 \right] = \Sigma_{-j,-j}^{-1} \Sigma_{-j,j} = -\Omega_{jj}^{-1} \Omega_{-j,j}.$$

Cycling through the variables, this approach can be used (Sun and Zhang, 2013) to construct an estimate of the entire matrix Ω. This is not always necessary, since entries of Ω encode

the partial correlations between pairs of variables (Whittaker, 1990):

$$\rho_{jk|-jk} = -\frac{\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}}, \tag{1.9}$$

so in the Gaussian case it is sufficient to estimate the support of Ω in order to identify the set of conditional independencies. This is a much more useful notion of dependence between variables than its marginal counterpart, upon which it is rarely reasonable to place a sparsity assumption. These conditional dependence relationships can be conveniently represented in the conditional independence graph, which is an undirected graph in which two nodes are connected by an edge if and only if they are conditionally dependent given the rest.

This nodewise approach towards graphical estimation (Meinshausen and Bühlmann, 2006; Sun and Zhang, 2013) is not without alternatives for estimation of a sparse inverse covariance (or its support) in this setting. A popular approach is the 'graphical Lasso' (Friedman et al., 2008; Yuan, 2010; Yuan and Lin, 2007), which constructs a single penalised objective for estimation of Ω under a sparsity assumption,

$$\hat{\Omega} \in \underset{\Omega \succeq 0}{\arg\min} \left\{ \operatorname{tr}(\hat{\Sigma}\Omega) - \log\det(\Omega) + \lambda \sum_{j \neq k} |\Omega_{jk}| \right\}. \tag{1.10}$$

Another alternative for estimation of the full precision matrix is the constrained ℓ_1 norm minimisation approach (CLIME) (Cai et al., 2011),

$$\hat{\Omega} \in \underset{\Omega \in \mathbb{R}^{p \times p}}{\arg\min} \|\Omega\|_1, \quad \text{subject to } \|\hat{\Sigma}\Omega - I_p\|_\infty \leq \lambda. \tag{1.11}$$

Estimation of the precision matrix has applications beyond the conditional independence graph: recall that we require such an estimate $\hat{\Omega}$ for construction of confidence intervals for coefficients in the high-dimensional linear model (1.5). Here, there is an important difference between estimation of the precision matrix, and construction of a surrogate for other statistical tasks. For this application, control of the quantity $\|\hat{\Omega}\hat{\Sigma} - I_p\|_\infty$ is required. Javanmard and Montanari (2014) choose to construct an estimate of Ω within the same constraint set as used in (1.11), and more recent related work (Yu et al., 2021) uses CLIME itself. Zhang and Zhang (2014) and van de Geer et al. (2014) both use $\hat{\Omega}$ constructed via the nodewise approach (1.8) in such a way that the KKT conditions for the regressions imply control over $\|\hat{\Omega}\hat{\Sigma} - I_p\|_\infty$.

1.2 Summary of chapters

The remainder of this thesis consists of three chapters, each investigating and developing methodology for solving high-dimensional statistical problems in a computationally efficient

way. The focus of the second chapter is in developing a novel approach to modelling categorical covariates. The third and fourth chapters consider variable screening from two very different perspectives, the former exploiting potential prior information to improve prediction performance and the latter speeding up neighbourhood selection in the presence of possible latent confounders.

Variable selection within a high-dimensional setting is challenging because the number of possible submodels grows exponentially in dimension. Often a problem will be endowed with some additional structure that informs the nature of the variable selection that is required. In the examples in Section 1.1.1 we have noted that previous off-the-shelf approaches were not suitable. The imposition of a particular choice of regulariser promotes sparsity in the solution in line with the beliefs of the practitioner, so an inappropriate choice will both worsen performance and impede the interpretability of the resulting model.

In Chapter 2 we consider the problem of modelling high-dimensional nominal categorical covariates. As discussed in Section 1.1.1, a key difficulty of this problem is that it doesn't involve a straightforward binary selection problem, with which we are relatively familiar, but instead a one-dimensional clustering problem within each of the variables. The number of such clusterings for a single variable with K categories is the K^{th} Bell number, a quantity that grows much faster than 2^K. As with other regression problems, a penalised least squares approach is attractive for reasons such as interpretability, but existing options for this approach involve penalty terms between each pair of coefficients within a variable, which is both methodologically undesirable and computationally expensive. Certain convex alternatives involving only a linear number of penalty terms are not suitable, as we show that they do not provide the desired fusion of levels. Approaches outside the penalised least squares framework exist, but can be impractical to compute, hard to interpret, and difficult to tune in a multivariate setting.

Our (necessarily) nonconvex approach retains the parsimony of successful procedures for other settings, requiring only as many penalty terms as there are coefficients. Its estimates enjoy low bias due to the properties of nonconvex penalties (Zhang and Zhang, 2012), and additionally in comparison to all-pairs type penalties.

The structure of the penalty permits the use of ideas from fast dynamic programming algorithms for solving the fused Lasso (Johnson, 2013), and its fast computation relies on a number of fortunate properties of the optimisation problem we define, such as a conservation of ordering of coefficients. This novel approach guarantees the global optimality of the solution after finitely many steps, despite the highly nonconvex nature of the problem. Implementationally, its intricate structure involves a number of nested functions, thus its efficiency is best exploited in a low-level programming language within minimal overhead for function calls and 'for loops'. For this reason the R package `CatReg` (Stokell et al., 2021a), which implements our algorithm, makes use of C++ routines for this computation.

Our main theoretical results concern the oracle least squares estimate, which is the ordinary least squares estimate making use of oracular knowledge of the true groupings. We establish sufficient conditions under which the oracle least squares estimate is a global optimum of the objective in the univariate setting; furthermore, the signal separation requirements cannot be relaxed for any estimator. This has a pleasing correspondence with our computational approach which solves this problem exactly, returning a global optimum which will be almost certainly unique. The multivariate counterpart of this result shows the oracle least squares estimate to be a fixed point of the blockwise coordinate descent routine we embed our algorithm within, with its proof technique mirroring our computational strategy. The proof of this result requires a novel approach, studying our objective and directly extracting properties of its optima until we have enough to specify sufficient conditions on the true coefficients in order for it to hold with high probability.

Our flexible and practical method, and its accompanying software, enables its use in a range of settings. We demonstrate this with a number of experiments on real, synthetic, and semi-synthetic datasets. A collection of thorough simulation studies establish the comptitive performance of our method across a range of low and high-dimensional settings. We also include as an illustration an error decomposition study of community models tracking the COVID-19 epidemic in the U.S. through 2020.

Variable screening is a particularly valuable string in the bow of variable selection methodology. The somewhat subtle distinction can be thought of as regarding the refinement and the focus of the approach. We see screening as primarily concerned with excluding large numbers of variables that we believe will not be informative in our modelling, a step which will typically be combined with a more refined variable selection and model fitting method.

Chapter 3 develops a practical and computationally efficient strategy to simultaneously fit generalised linear models and perform variable screening based on possible prior information on the importance of variables. We propose a simple system of fitting a sequence of models over nested subsets of variables with respect to the given ordering over variables. This enables a model to potentially be fitted on a much smaller set of variables in the case where this ordering is highly informative, with a commensurate improvement in the model's prediction performance. If, on the other hand, the information is not helpful, there is only a small cost associated with using this approach compared to the fitting of a single model over the full set of variables.

While the computational cost of naïvely fitting all of these models may be far too high, this can often be avoided. For Lasso and ridge regression models, arguably cornerstones of high-dimensional regression methodology, we provide and implement computational strategies to vastly reduce the cost of our approach. For ridge regression, this allows for a potentially very large number of possible models to be compared, with an average cost of just $O(n^2)$. Experiments show that in practice our method is fast and flexible, while enabling noticeable

improvements in performance. We establish that even in cases where the ordering is only slightly more useful than a random permutation, the benefits from this method can be enough to warrant its use. We conclude with discussion of an exciting direction for the future extension of this approach for use with potential prior information in more general forms than orderings, providing a framework for using our computational strategy within a nested tree structure.

In the Chapter 4 we consider a different variable screening problem, here from the perspective of reducing the computational cost of identifying the dependence structure of high-dimensional data. We focus on the setting of conditional independence graph estimation within Gaussian graphical models, which are a popular model for many high-dimensional settings such as in finance or gene expression data. We argue that in this setting, one should avoid assuming a minimal marginal correlation between any pair of variables connected by an edge. These relations are often given causal interpretation via a directed graphical model, from which it becomes clear that such an assumption will not be satisfied for a wide range of graphical structures. We develop a two-stage screening procedure and show that this restrictive assumption can be avoided, while still returning sufficiently small screening sets which are fast to compute in practice. Under realistic assumptions we show that our approach will include all true edges with high probability. Furthermore, avoiding the need for multiple testing corrections via the strategy we use for computing test statistics implicitly increases its sensitivity.

An additional barrier to the estimation of conditional independence graphs (and any variable screening beforehand) is the presence of latent confounding. Our approach combines screening and estimation in a single procedure, only requiring a covariance estimate from the data. As a result, it can be made robust to confounding as straightforwardly as by plugging in a covariance estimate with this property. In particular, its scale invariance in the covariance estimate enables us to make use of recent developments in the covariance estimation literature which can recover the covariance in the latent confounding setting, up to an unknown constant. A range of experiments show that our combined screening and estimation approach performs well across a range of settings, in particular those with unobserved confounder variables.

1.3 Notation

For much of thesis, we will use the following notational conventions. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, we denote its i^{th} row $\boldsymbol{x}_i \in \mathbb{R}^p$ and its j^{th} column $\boldsymbol{X}_j \in \mathbb{R}^n$. We use the subscript '$-j$' to denote all columns except j, e.g. $\boldsymbol{X}_{-j} \in \mathbb{R}^{n \times (p-1)}$ is the submatrix of $\boldsymbol{X}$ with the j^{th} column removed.

Chapter 2

Modelling high-dimensional categorical covariates

2.1 Introduction

Categorical data arise in a number of application areas. For example, electronic health data typically contain records of diagnoses received by patients coded within controlled vocabularies and also prescriptions, both of which give rise to categorical variables with large numbers of levels (Jensen et al., 2012). Vehicle insurance claim data also contain a large number of categorical variables detailing properties of the vehicles and parties involved (Hu et al., 2018). When performing regression with such data as covariates, it is often helpful, both for improved predictive performance and interpretation of the fit, to fuse the levels of several categories together in the sense that the estimated coefficients corresponding to these levels have exactly the same value.

To fix ideas, consider the following ANOVA model relating response vector $\boldsymbol{Y} = [Y_1, \ldots, Y_n]^T \in \mathbb{R}^n$ to categorical predictors $X_{ij} \in \{1, \ldots, K_j\}$, $j = 1, \ldots, p$:

$$Y_i = \mu^0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk}^0 \mathbb{1}_{\{X_{ij}=k\}} + \varepsilon_i. \tag{2.1}$$

Here the ε_i are independent zero mean random errors, μ^0 is a global intercept and θ_{jk}^0 is the contribution to the response of the k^{th} level of the j^{th} predictor; we will later place restrictions on the parameters to ensure they are identifiable. We are interested in the setting where the coefficients corresponding to any given predictor are clustered, so defining

$$s_j = |\{\theta_{j1}^0, \ldots, \theta_{jK_j}^0\}|, \tag{2.2}$$

we have $s_j \ll K_j$, at least when K_j is large. Note that our setup can include high-dimensional settings where p is large and many of the predictors do not contribute at all to the response: when $s_j = 1$, the contribution of the j^{th} predictor is effectively null as it may be absorbed by the intercept term.

2.1.1 Background and motivation

Early work on collapsing levels together in low-dimensional models of the form (2.1) focused on performing a variety of significance tests for whether certain sets of parameters were equal (Calinski and Corsten, 1985; Scott and Knott, 1974; Tukey, 1949). A more modern and algorithmic method based on these ideas is Delete or merge regressors (Maj-Kańska et al., 2015), which involves agglomerative clustering based on t-statistics for differences between levels.

The CART algorithm (Breiman et al., 1984) for building decision trees effectively starts with all levels of the variables fused together and greedily selects which levels to split. One potential drawback of these greedy approaches is that in high-dimensional settings where the search space is very large, they may fail to find good groupings of the levels. The popular random forest procedure (Breiman, 2001) uses randomisation to alleviate the issues with the greedy nature of the splits, but sacrifices interpretability of the fitted model.

An alternative to greedy approaches in high-dimensional settings is using penalty-based methods such as the Lasso (Tibshirani, 1996). This can be applied to continuous or binary data and involves optimising an objective for which global minimisation is computationally tractable, thereby avoiding some of the pitfalls of greedy optimisation. In contrast to random forest, the fitted models are sparse and interpretable. Inspired by the success of the Lasso and related methods for high-dimensional regression, a variety of approaches have proposed estimating $\boldsymbol{\theta}^0 = (\theta^0_{jk})_{j=1,\ldots,p,\, k=1,\ldots,K_j}$ and μ_0 via optimising over $(\mu, \boldsymbol{\theta})$ a sum of a least squares criterion

$$\ell(\mu, \boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \mu - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}_{\{X_{ij}=k\}} \right)^2 \tag{2.3}$$

and a penalty of the form

$$\sum_{j=1}^{p} \sum_{k=2}^{K_j} \sum_{l=1}^{k-1} w_{j,kl} |\theta_{jk} - \theta_{jl}| \,. \tag{2.4}$$

This is the CAS-ANOVA penalty of Bondell and Reich (2009). The weights $w_{j,kl}$ can be chosen to balance the effects of having certain levels of categories more prevalent than others in the data. The penalty is an 'all-pairs' version of the fused Lasso and closely related to so-called convex clustering (Chiquet et al., 2017; Hocking et al., 2011). We note that there are several other approaches besides using penalty functions. For instance, Pauger and Wagner (2019) proposes a Bayesian modelling procedure using sparsity-inducing prior distributions

to encourage fusion of levels. See also Tutz and Gertheiss (2016) and references therein for a review of other methods including those based on mixture models and kernels.

The fact that the optimisation problem resulting from (2.4) is convex makes the procedure attractive. However, a drawback is that it may not give a desirable form of shrinkage. Indeed, consider the case where $p = 1$, and dropping subscripts for simplicity, all $w_{kl} = 1$. This would typically be the case if all levels were equally prevalent. Further suppose for simplicity that the number of levels K is even. Then if the coefficients are clustered into two groups where one contains only a single isolated coefficient, the number of non-zero summands in (2.4) is only $K - 1$. This almost doubles to $2(K - 2)$ when one of the two groups is of size 2. The extreme case where the two groups are of equal size yields $(K/2)^2$ non-zero summands. This particular property of all-pairs penalties, which results in them favouring groups of unequal sizes, is illustrated schematically in Figure 2.1. We can see the impact of this in the following

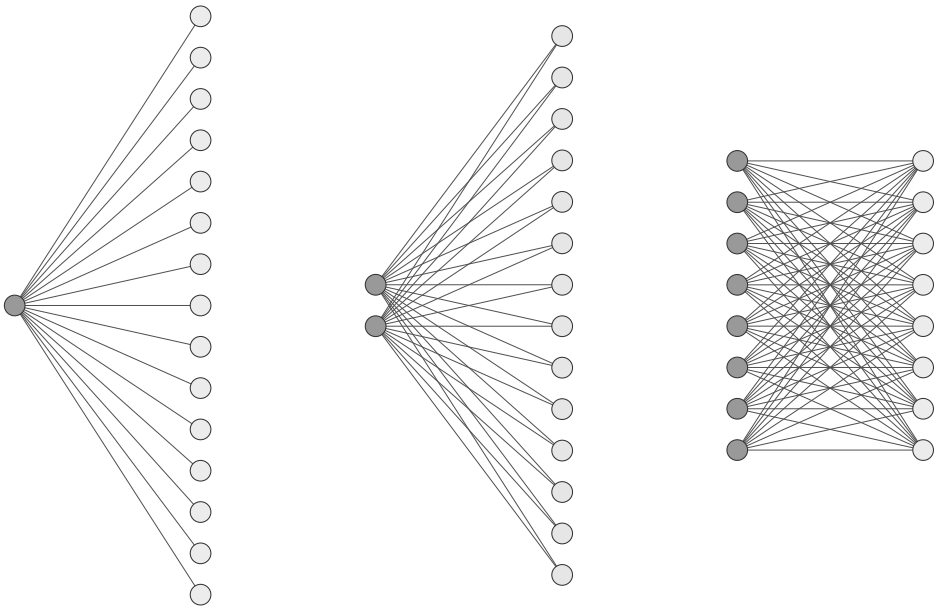


Figure 2.1 Illustration of the number of non-zero summands in (2.4) when $p = 1$, $K = 16$ and coefficients are clustered into two groups of equal size (right), and where one contains a single coefficient (left) and two coefficients (middle).

concrete example. Suppose $K = 20$ levels are clustered into four groups with

$$\theta_1^0 = \cdots = \theta_4^0 = -6, \qquad \theta_5^0 = \cdots = \theta_{10}^0 = -2.5$$
$$\theta_{11}^0 = \cdots = \theta_{16}^0 = 2.5, \qquad \theta_{17}^0 = \cdots = \theta_{20}^0 = 6.$$

If the coefficient estimates satisfy $\hat{\theta}_1 = \cdots = \hat{\theta}_4 < \hat{\theta}_5 = \cdots = \hat{\theta}_{10} \leq \hat{\theta}_k$ for all $k \geq 11$, so the first two groups have distinct coefficients, then moving any coefficient from the first group towards the second, and so increasing the number of estimated groups, actually *decreases* the

penalty contribution in (2.4). Specifically, if the k^{th} coefficient for some $k \in \{1, \ldots, 4\}$ moves to $\hat{\theta}_k + t$ for $t \in [0, \hat{\theta}_5 - \hat{\theta}_4]$ with all other coefficients kept fixed, the penalty contribution decreases by $13t$. In this case then, CAS-ANOVA will struggle to keep the groups intact, especially smaller ones. We see this in Figure 2.2, which shows the result of applying CAS-ANOVA to data generated according to (2.1) with $p = 1$, $\boldsymbol{\theta}^0$ as above, $n = 20$ (so we have a single observation corresponding to each level), and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. There is no value of the tuning parameter λ where the true groups are recovered.

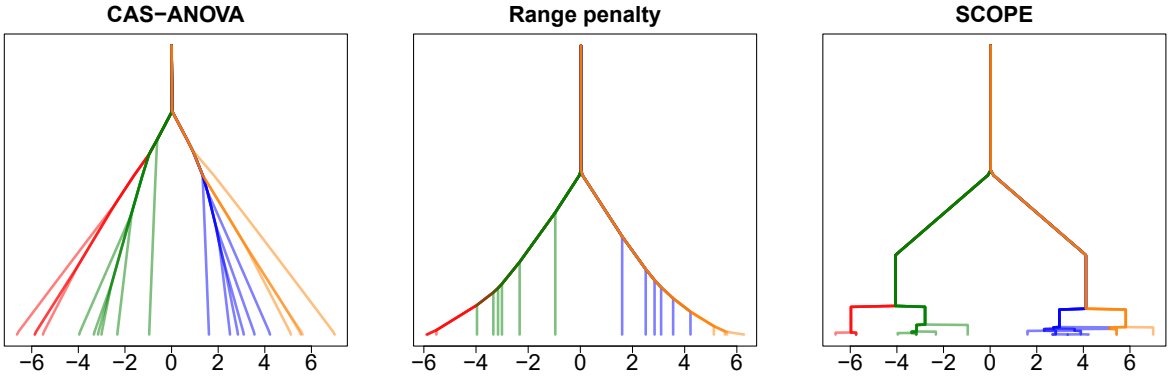


Figure 2.2 Solution paths as the tuning parameter varies in a univariate example where there are four true groups. From left to right: CAS-ANOVA, the range penalty and SCOPE with $\gamma = 8$. The setup is as described in the main text of Section 2.1.1, with the different colours corresponding to the different true groups. The tuning parameter varies along the y axis. In this example, only SCOPE identifies the 4 correct groups at any point along its solution path.

As in the standard regression setting, the bias introduced by all-pairs ℓ_1-type penalties may be reduced by choosing data-adaptive weights analogously to the adaptive Lasso (Zou, 2006), or replacing the absolute value $|\theta_{jk} - \theta_{jl}|$ by $\rho(|\theta_{jk} - \theta_{jl}|)$ where ρ is a concave and non-decreasing penalty function (Ma and Huang, 2017; Oelker et al., 2015). However, this does not address the basic issue of a preference for groups of unequal sizes. Additionally, optimising an objective involving a penalty with $O\left(\sum_{j=1}^{p} K_j^2\right)$ summands can be computationally challenging, particularly in the case where ρ is not convex, both in terms of runtime and memory.

To help motivate the new approach we are proposing in this chapter, let us consider the setting where the predictors are ordinal rather than nominal, so there is an obvious ordering among the levels. In these settings, it is natural to consider a fused Lasso (Tibshirani et al., 2005) penalty of the form

$$\sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} |\theta_{j\pi_j(k+1)} - \theta_{j\pi_j(k)}|, \tag{2.5}$$

where π_j is a permutation of $\{1, \ldots, K_j\}$ specifying the given order; this is done in Gertheiss and Tutz (2010) who advocate using it conjunction with the all-pairs-type CAS-ANOVA penalty for nominal categories.

If however we treat the nominal variable setting as analogous to having ordinal variables with unknown orderings π_j, one might initially think of choosing π_j corresponding to the order of the estimates $\boldsymbol{\theta}_j = (\theta_{jk})_{k=1}^{K_j}$, such that $\theta_{j\pi_j(k)} = \theta_{j(k)}$, where $\theta_{j(k)}$ is the k^{th} smallest entry in $\boldsymbol{\theta}_j$. This however leads to what we refer to as the 'range' penalty:

$$\sum_{k=1}^{K_j-1} |\theta_{j(k+1)} - \theta_{j(k)}| = \max_k \theta_{jk} - \min_k \theta_{jk}. \tag{2.6}$$

Whilst this shrinks the largest and smallest of the estimated coefficients together, the remaining coefficients lying in the open interval between these are unpenalised and so no grouping of the estimates is encouraged, as we observe in Figure 2.2; see also Oelker et al. (2015) for a discussion of this issue in the context of ordinal variables.

2.1.2 Our contributions and organisation of the chapter

Given how all-pairs penalties have an intrinsic and undesirable preference for unequal group sizes, and how the fused Lasso applied to ordered coefficients (2.6) does not result in grouping of the coefficients, we propose the following solution. Our approach is to use the penalty

$$\sum_{j=1}^{p} \sum_{k=1}^{K_j-1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}),$$

for concave (and nonconvex) non-decreasing penalty functions ρ_j, which, for computational reasons we discuss in Section 2.3, we base on the minimax concave penalty (MCP) (Zhang, 2010). In Section 2.2 we formally introduce our method, which we call SCOPE, standing for **S**parse **C**oncave **O**rdering & **P**enalisation **E**stimator.

Note that whereas in conventional high-dimensional regression, the use of nonconvex penalties has been primarily motivated by a need to reduce bias in the estimation of large coefficients (Fan and Li, 2001), here the purpose is very different: in our setting a nonconvex penalty is in fact even necessary for shrinkage to sparse solutions to occur (see Proposition 2.1). Because of these fundamental differences, the rich algorithmic and statistical theory concerning high-dimensional regression with nonconvex penalties (see for example Fan et al. (2018); Loh and Wainwright (2012, 2015); Wang et al. (2014); Zhao et al. (2018) and references therein) is not directly applicable to our setting.

In Section 2.3, we therefore introduce a new dynamic programming approach that recovers the global minimum of the resulting objective function exactly in the univariate case, i.e. when $p = 1$. We then build this into a blockwise coordinate descent approach to tackle the multivariate setting.

In Section 2.4 we study the theoretical properties of SCOPE and give sufficient conditions for the estimator to coincide with the least squares solution with oracular knowledge of the

level fusions in the univariate case. These conditions involve a minimal separation between unequal coefficients that is, up to constant factors, minimax optimal. Our result contrasts sharply with Theorem 2 of Ma and Huang (2017) for an all-pairs nonconvex penalty. The latter instead shows the existence of a local optimum that coincides with the oracle least squares solution. Whilst in conventional high-dimensional regression settings, it is known that under certain conditions, all local optima have favourable properties (Loh and Wainwright, 2015), we note that the separation requirements in Ma and Huang (2017) are substantially weaker than those indicated by the minimax lower bound, and so cannot be extended to a particular local optimum determined by the data; see the discussion following Theorem 2.5.

We use our univariate result to show that the oracle least squares solution is a fixed point of our blockwise coordinate descent algorithm in the multivariate case. In Section 2.5 we outline some extensions of our methodology including a scheme for handling settings when there is a hierarchy among the categorical variables. Section 2.6 contains numerical experiments that demonstrate the favourable performance of our method compared to a range of competitors on both simulated and real data. We conclude with a discussion in Section 2.7. Sections 2.9 and 2.10 contain all of the proofs. Further details of our algorithm can be found in Section 2.8, as well as additional information on the runtime of our algorithm and an approximate version suitable for very large-scale settings.

2.2 Methodology

Recall that our goal is to estimate parameters $(\mu^0, \boldsymbol{\theta}^0)$ in model (2.1). Let us first consolidate some notation. For any $\boldsymbol{\theta} = (\theta_{jk})_{j=1,\ldots,p,\, k=1,\ldots,K_j} \in \mathbb{R}^{K_1} \times \cdots \times \mathbb{R}^{K_p}$, we define $\boldsymbol{\theta}_j = (\theta_{jk})_{k=1}^{K_j} \in \mathbb{R}^{K_j}$. We will study the univariate setting where $p = 1$ separately, and so it will be helpful to introduce some simplified notation for this case, dropping any extraneous subscripts. We thus write $K \equiv K_1$, $X_i \equiv X_{i1}$ and $\rho \equiv \rho_1$. Additionally, we let $\bar{Y}_k$ denote the average of the Y_i with $X_i = k$:

$$\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n} Y_i \mathbb{1}_{\{X_i = k\}}, \tag{2.7}$$

where $n_k = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = k\}}$, and $\bar{\boldsymbol{Y}} = (\bar{Y}_k)_{k=1}^{K} \in \mathbb{R}^K$.

In order to avoid an arbitrary choice of corner point constraint, we instead impose the following to ensure that $\boldsymbol{\theta}^0$ is identifiable: for all $j = 1, \ldots, p$ we have

$$g_j(\boldsymbol{\theta}_j^0) = 0, \text{ where } g_j(\boldsymbol{\theta}_j) = \sum_{k=1}^{K_j} n_{jk} \theta_{jk} \text{ and } n_{jk} = \sum_{i=1}^{n} \mathbb{1}_{\{X_{ij} = k\}}. \tag{2.8}$$

We note that this requirement is solely for modelling purposes and does not impose any restriction on the data generating mechanism itself. In a random design setting, the coefficient vector can be reparameterised in a way that is consistent with the constraint (2.8). In the

absence of any constraint, the model would not be identifiable due the overspecification of the intercept.

Let $\Theta_j = \{\boldsymbol{\theta}_j \in \mathbb{R}^{K_j} : g_j(\boldsymbol{\theta}_j) = 0\}$, and let $\Theta = \Theta_1 \times \cdots \times \Theta_p$. We will construct estimators by minimising over $\mu \in \mathbb{R}$ and $\boldsymbol{\theta} \in \Theta$ an objective function of the form

$$\tilde{Q}(\mu, \boldsymbol{\theta}) = \ell(\mu, \boldsymbol{\theta}) + \sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}),$$

where ℓ is the least squares loss function (2.3) and $\theta_{j(1)} \leq \cdots \leq \theta_{j(K_j)}$ are the order statistics of $\boldsymbol{\theta}_j$. We allow for different penalty functions ρ_j for each predictor in order to help balance the effects of varying numbers of levels K_j. The identifiability constraint that $\boldsymbol{\theta} \in \Theta$ ensures that the estimated intercept $\hat{\mu} = \arg\min_\mu \tilde{Q}(\mu, \boldsymbol{\theta})$ satisfies $\hat{\mu} = \sum_{i=1}^n Y_i/n$.

We note that whilst the form of the identifiability constraint would not have a bearing on the fitted values of unregularised least squares regression, this is not necessarily the case when regularisation is imposed. For example, consider the simple univariate setting with $p = 1$ and the corner point constraint $\theta_1 = 0$. Then the fitted value for an observation with level 1 would simply be the average $\bar{Y}_1$, coinciding with that of unpenalised least squares. However the fitted values with observations with other level $k \geq 2$ would be subject to regularisation and in general be different to $\bar{Y}_k$. This inequitable treatment of the levels is clearly undesirable as they may have been labelled in an arbitrary way. Our identifiability constraint treats the levels more symmetrically, but also takes into account the prevalence of levels, so the fitted values corresponding to more prevalent levels effectively undergo less regularisation.

As the estimated intercept $\hat{\mu}$ does not depend on the tuning parameters, we define

$$Q(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \hat{\mu} - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}_{\{X_{ij}=k\}} \right)^2 + \sum_{j=1}^{p} \sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}). \tag{2.9}$$

We will take the regularisers $\rho_j : [0, \infty) \to [0, \infty)$ in (2.9) to be concave (and nonconvex); as discussed in the introduction and formalised in Proposition 2.1 below, a nonconvex penalty is necessary for fusion to occur.

Proposition 2.1. *Consider the univariate case with $p = 1$. Suppose the subaverages $(\bar{Y}_k)_{k=1}^K$ (2.7) are all distinct, and that $\rho_1 \equiv \rho$ is convex. Then any minimiser $\hat{\boldsymbol{\theta}}$ of Q has $\hat{\theta}_k \neq \hat{\theta}_l$ for all $k \neq l$ such that $\hat{\theta}_{(1)} < \bar{Y}_k - \hat{\mu} < \hat{\theta}_{(K)}$ or $\hat{\theta}_{(1)} < \bar{Y}_l - \hat{\mu} < \hat{\theta}_{(K)}$.*

This result states that with a convex choice for ρ_j, no coefficients will be fused unless they take the minimal or maximal values within that variable. This general behaviour is particularly intuitive for the range penalty (2.6) example, illustrated in Figure 2.2. A strictly nonconvex function is therefore necessary, and we base the penalties $\rho_j : [0, \infty) \to [0, \infty)$ on

the MCP (Minimax Concave Penalty) (Zhang, 2010):

$$\rho(x) = \rho_{\gamma,\lambda}(x) = \int_0^x \lambda \left(1 - \frac{t}{\gamma\lambda}\right)_+ dt,$$

where $(u)_+ = u\mathbb{1}_{\{u \geq 0\}}$. This is a piecewise quadratic function with gradient λ at 0 and flat beyond $\gamma\lambda$. For computational reasons which we discuss in Section 2.3, the simple piecewise quadratic form of this is particularly helpful. In the multivariate case we take $\rho_j = \rho_{\gamma,\lambda_j}$ with $\lambda_j = \lambda\sqrt{K_j}$. This choice of scaling is motivated by requiring that when $\boldsymbol{\theta}^0 = 0$ we also have $\hat{\boldsymbol{\theta}} = 0$ with high probability; see Lemma 2.10 in Section 2.10.1. We discuss the choice of the tuning parameters λ and γ in Section 2.3.3, but first turn to the problem of optimising (2.9).

2.3 Computation

In this section we include details of how SCOPE is computed. Section 2.3.1 motivates and describes the dynamic programming algorithm we use to compute global minimiser of the SCOPE objective, which is highly non-convex. Section 2.3.2 contains details of how this is used to solve the multivariate objective by embedding it within a blockwise coordinate descent routine. Discussion of practical considerations is contained in Section 2.3.3.

2.3.1 Univariate model

Preliminaries

We now consider the univariate case ($p = 1$) and explain how the solutions are computed. In this case, we may rewrite the least squares loss contribution to the objective function in the following way.

$$\begin{aligned}\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu} - \sum_{k=1}^{K}\theta_k\mathbb{1}_{\{X_i=k\}}\right)^2 &= \frac{1}{2n}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{1}_{\{X_i=k\}}(Y_i - \hat{\mu} - \theta_k)^2 \\ &= \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \hat{\mu} - \theta_k)^2 + \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{1}_{\{X_j=k\}}(Y_i - \bar{Y}_k)^2\end{aligned} \tag{2.10}$$

where $w_k = n_k/n$. Thus the optimisation problem (2.9) can be written equivalently as

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta}\left\{\frac{1}{2}\sum_{k=1}^{K} w_k\left(\bar{Y}_k - \hat{\mu} - \theta_k\right)^2 + \sum_{k=1}^{K-1}\rho\left(\theta_{(k+1)} - \theta_{(k)}\right)\right\}, \tag{2.11}$$

suppressing the dependence of the MCP ρ on tuning parameters γ and λ. In fact, it is straightforward to see that the constraint that the solution lies in Θ will be automatically satisfied, so we may replace Θ with $\mathbb{R}^K$. Two challenging aspects of the optimisation problem

above are the presence of the nonconvex ρ and the order statistics. The latter however are easily dealt with using the result below, which holds more generally whenever ρ is a concave function.

Proposition 2.2. *Consider the univariate optimisation* (2.11) *with ρ any concave function such that a minimiser $\hat{\boldsymbol{\theta}}$ exists. If for k, l we have $\bar{Y}_k > \bar{Y}_l$, then $\hat{\theta}_k \geq \hat{\theta}_l$.*

This observation substantially simplifies the optimisation: after re-indexing such that $\bar{Y}_1 \leq \bar{Y}_2 \leq \cdots \leq \bar{Y}_K$, we may re-express (2.11) as,

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}:\theta_1 \leq \cdots \leq \theta_K} \left\{ \frac{1}{2} \sum_{k=1}^{K} w_k \left(\bar{Y}_k - \hat{\mu} - \theta_k \right)^2 + \sum_{k=1}^{K-1} \rho\left(\theta_{k+1} - \theta_k\right) \right\}. \tag{2.12}$$

We use the following intermediate functions to structure the algorithm:

$$\begin{aligned} f_1(\theta_1) &= \frac{1}{2} w_1 (\bar{Y}_1 - \hat{\mu} - \theta_1)^2, \\ f_k(\theta_k) &= \min_{\theta_{k-1}:\theta_{k-1} \leq \theta_k} \{ f_{k-1}(\theta_{k-1}) + \rho(\theta_k - \theta_{k-1}) \} + \frac{1}{2} w_k (\bar{Y}_k - \hat{\mu} - \theta_k)^2, \\ b_k(\theta_k) &= \operatorname*{sarg\,min}_{\theta_{k-1}:\theta_{k-1} \leq \theta_k} \{ f_{k-1}(\theta_{k-1}) + \rho(\theta_k - \theta_{k-1}) \}, \end{aligned} \tag{2.13}$$

for $k = 2, \ldots, K$; here sarg min refers to the smallest minimiser in the case that it is not unique. Invariably however this will be unique, as the following result indicates.

Proposition 2.3. *The set of $(\bar{Y}_k)_{k=1}^K$ that yields distinct solutions to* (2.11) *has Lebesgue measure zero as a subset of $\mathbb{R}^K$.*

We will thus tacitly assume uniqueness in some of the discussion that follows, though this is not required for our algorithm to return a global minimiser. Observe now that $\hat{\theta}_K$ is the minimiser of the univariate objective function f_K: indeed for $k \geq 2$,

$$f_k(\theta_k) = \min_{(\theta_1, \ldots, \theta_{k-1}):\theta_1 \leq \cdots \leq \theta_{k-1} \leq \theta_k} \left\{ \frac{1}{2} \sum_{l=1}^{k} w_l (\bar{Y}_l - \hat{\mu} - \theta_l)^2 + \sum_{l=1}^{k-1} \rho(\theta_{l+1} - \theta_l) \right\}. \tag{2.14}$$

Furthermore, we have $\hat{\theta}_{K-1} = b_K(\hat{\theta}_K)$, and more generally $\hat{\theta}_k = b_{k+1}(\hat{\theta}_{k+1})$ for $k = K - 1, \ldots, 1$. Thus provided f_K can be minimised efficiently (which we shall see is indeed the case), given this and the functions $b_2, \ldots, b_K$ we can iteratively compute $\hat{\theta}_K, \hat{\theta}_{K-1}, \ldots, \hat{\theta}_1$. In order to make use of these properties, we must be able to compute f_K and the b_k efficiently; we explain how to do this in the following subsection.

Computation of f_K and $b_2, \ldots, b_K$

The simple piecewise quadratic form of the MCP-based penalty is crucial to our approach for computing the f_K and the b_k. Some important consequences of this piecewise quadratic property are summarised in the following lemma.

Lemma 2.4. *For each k,*

(i) f_k is continuous, coercive and piecewise quadratic with finitely many pieces;

(ii) b_k is piecewise linear with finitely many pieces;

(iii) for each $\theta_{k+1} \in \mathbb{R}$, if a minimiser $\tilde{\theta}_k = \tilde{\theta}_k(\theta_{k+1})$ of $\theta_k \mapsto f_k(\theta_k) + \rho(\theta_{k+1} - \theta_k)$ over $(-\infty, \theta_{k+1}]$ satisfies $\tilde{\theta}_k < \theta_{k+1}$, then f_k must be differentiable at $\tilde{\theta}_k$.

Properties (i) and (ii) above permit exact representation of f_k and b_k with finitely many quantities. The key task then is to form the collection of intervals and corresponding coefficients of quadratic functions for

$$g_k(\theta_{k+1}) = \min_{\theta_k : \theta_k \leq \theta_{k+1}} \{f_k(\theta_k) + \rho(\theta_{k+1} - \theta_k)\} \tag{2.15}$$

given a similar piecewise quadratic representation of f_k; and also the same for the linear functions composing b_k. A piecewise quadratic representation of f_{k+1} would then be straightforward to compute, and we can iterate this process. To take advantage of property (iii) above, in computing $g_k(\theta_{k+1})$ we can separately search for minimisers at stationary points in $(-\infty, \theta_{k+1})$ and compare the corresponding function values with $f_k(\theta_{k+1})$; the fact that we need only consider potential minimisers at points of differentiability will simplify things as we shall see below.

Suppose $I_{k,1}, \ldots, I_{k,m(k)}$ are intervals that partition $\mathbb{R}$ (closed on the left) and $q_{k,1}, \ldots, q_{k,m(k)}$ are corresponding quadratic functions such that $f_k(\theta_k) = q_{k,r}(\theta_k)$ for $\theta_k \in I_{k,r}$. Let us write

$$\tilde{q}_{k,r}(\theta_k) = \begin{cases} q_{k,r}(\theta_k) & \text{if } \theta_k \in I_{k,r} \\ \infty & \text{otherwise.} \end{cases}$$

We may then express f_k as $f_k(\theta_k) = \min_r \tilde{q}_{k,r}(\theta_k)$. We can also express the penalty $\rho = \rho_{\gamma,\lambda}$ in a similar fashion. Let

$$\tilde{\rho}_1(x) = -\gamma\lambda^2\{1 - x/(\gamma\lambda)\}^2/2 + \gamma\lambda^2/2 \text{ if } 0 \leq x < \gamma\lambda \text{ and } \infty \text{ otherwise,}$$
$$\tilde{\rho}_2(x) = \gamma\lambda^2/2 \text{ if } x \geq \gamma\lambda \text{ and } \infty \text{ otherwise.}$$

Then $\rho(x) = \min_t \tilde{\rho}_t(x)$ for $x \geq 0$. Let D_k be the set of points at which f_k is differentiable. We then have, using Lemma 2.4 (iii) that

$$\begin{aligned} g_k(\theta_{k+1}) &= \min_{\theta_k : \theta_k \leq \theta_{k+1}} \{\min_r \tilde{q}_{k,r}(\theta_k) + \min_t \tilde{\rho}_t(\theta_{k+1} - \theta_k)\} \\ &= \min[\tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \min_{r,t} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}, f_k(\theta_{k+1})] \\ &= \min[\min_{r,t} \tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}, f_k(\theta_{k+1})], \end{aligned} \tag{2.16}$$

where $\tilde{\min}$ denotes the minimum if it exists and ∞ otherwise. The fact that in the inner minimisation we are permitted to consider only points in D_k simplifies the form of

$$u_{k,r,t}(\theta_{k+1}) = \tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}. \tag{2.17}$$

We show in Section 2.8.1 that this is finite only on an interval and there takes the value of a quadratic function; coefficients for this function and the interval endpoints have closed form expressions that are elementary functions of the coefficients and intervals corresponding to $\tilde{q}_{k,r}$. With this, we have an explicit representation of g_k as the minimum of a collection of functions that are quadratic on intervals and ∞ everywhere else. Let us refer to these intervals (closed on the left) and corresponding quadratic functions as $J_{k,1}, \ldots, J_{k,n(k)}$ and $p_{k,1}, \ldots, p_{k,n(k)}$ respectively.

In order to produce a representation of f_{k+1} for use in future iterations, we must express g_k as a collection of quadratics defined on *disjoint* intervals. To this end, define for each $x \in \mathbb{R}$ the *active set at* x, $A(x) = \{r : x \in J_{k,r}\}$. Note that the endpoints of the intervals $J_{k,r}$ are the points where the active set changes and it is thus straightforward to determine $A(x)$ at each x. Let $r(x)$ be the index such that $g_k(x) = p_{k,r(x)}(x)$. For large negative values of x, $A(x)$ will contain a single index and for such x this must be $r(x)$. Consider also for each $r \in A(x) \setminus \{r(x)\}$, the horizontal coordinate x' of the first intersection beyond x (if it exists) between $p_{k,r}$ and $p_{k,r(x)}$. We refer to the collection of all such tuples (x', r) as the *intersection set at* x and denote it by $N(x)$. Given $r(x)$, $N(x)$ can be computed easily. The intersection set $N(x)$ then in turn helps to determine the smallest $x' > x$ where $r(x') \neq r(x)$ changes, that is the next knot of g_k beyond x, as we now explain. Suppose at a point x_{old}, we have computed $r_{\text{old}} = r(x_{\text{old}})$. We set $x_{\text{cur}} = x_{\text{old}}$ and perform the following.

1. Given $r(x_{\text{cur}})$, compute $N(x_{\text{cur}})$ and set $(x_{\text{int}}, r_{\text{int}}) = \arg\min_{(x,r) \in N(x_{\text{cur}})} x$.

2. If there are no changes in the active set between x_{cur} and x_{int}, we have found the next knot point at x_{int} and $r_{\text{int}} = r(x_{\text{int}})$.

3. If instead the active set changes, move x_{cur} to the leftmost change point. We have that $r(x) = r_{\text{old}}$ for $x \in [x_{\text{old}}, x_{\text{cur}})$. To determine if $r(x)$ changes at x_{cur}, we check if

(i) r_{old} leaves the active set at x_{cur}, so $r_{\text{old}} \notin A(x_{\text{cur}})$, or

(ii) some r_{new} enters the active set at x_{cur} and 'beats' r_{old}, so $r_{\text{new}} \in A(x_{\text{cur}}) \setminus A(x_{\text{old}})$ and $p_{k,r_{\text{new}}}(x_{\text{cur}} + \epsilon) < p_{k,r_{\text{old}}}(x_{\text{cur}} + \epsilon)$ for $\epsilon > 0$ sufficiently small.

If either hold x_{cur} is a knot and $r(x_{\text{cur}})$ may be computed via $r(x_{\text{cur}}) = \arg\min_{r \in A(x_{\text{cur}})} p_{k,r}(x_{\text{cur}})$. If neither hold, we conclude that $r(x_{\text{cur}}) = r_{\text{old}}$ and go to step 1 once more.

Hence we can proceed from one knot of g_k to the next by comparing the values and intersections of a small collection of quadratic functions, and thereby form a piecewise quadratic representation of g_k in a finite number of steps. Figure 2.3 illustrates the steps outlined above. The pieces of b_k may be computed in a similar fashion.

We note there are several modifications that can speed up the algorithm: for example, for each r, $u_{k,r,2}$ (2.17) is a constant function where it is finite (see $p_{k,2}$ in the figure), and these can be dealt with more efficiently. For further details including pseudocode see Section 2.8.2.

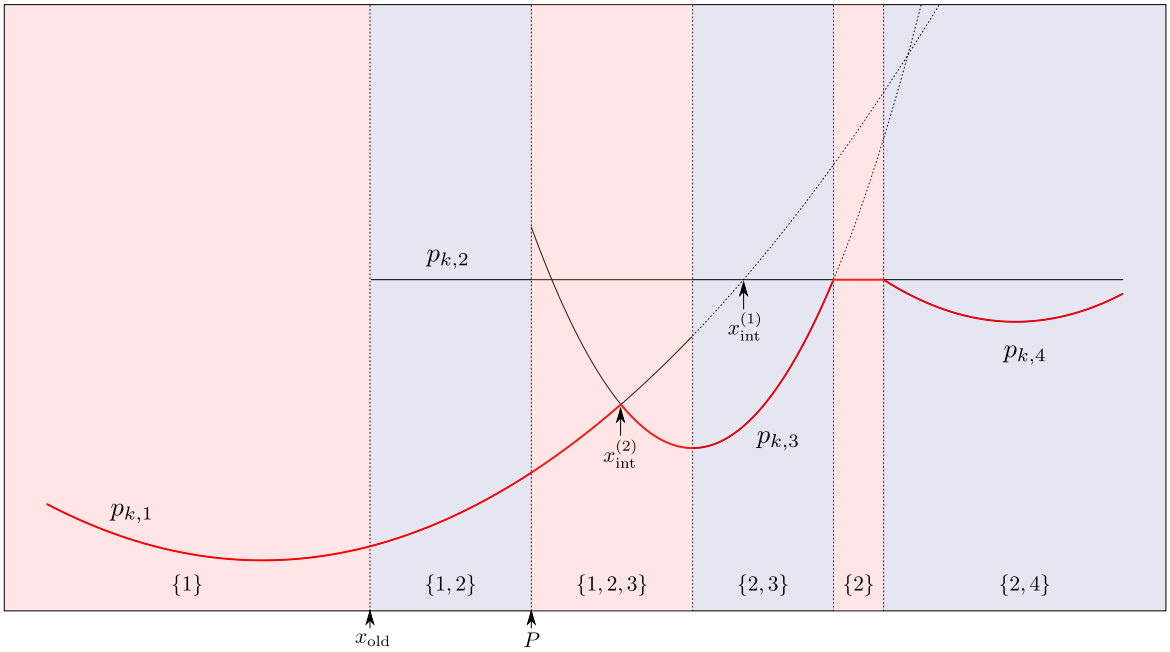


Figure 2.3 Illustration of the optimisation problem and our algorithm, to be interpreted with reference to steps 1, 2, 3 in the main text. Shading indicates regions where the active set, displayed at the bottom of the plot, is invariant, and vertical dotted lines signify changes. Dotted curves correspond to parts of quadratic functions $p_{k,l}$ lying outside their associated intervals $J_{k,l}$. At x_{old}, we have $r(x_{\text{old}}) = 1$, $A(x_{\text{old}}) = \{1, 2\}$ and $N(x_{\text{old}}) = \{(x_{\text{int}}^{(1)}, 2)\}$. Since the active set changes between x_{old} and $x_{\text{int}}^{(1)}$, we move x_{cur} to the first change point P and see neither (i) nor (ii) occur. We therefore return to step 1 and compute $N(x_{\text{cur}})$ which additionally contains $(x_{\text{int}}^{(2)}, 2)$. As the active set is unchanged between x_{cur} and $x_{\text{int}}^{(2)}$, we have determined the next knot point $x_{\text{int}}^{(2)}$ and minimising quadratic $p_{k,3}$.

In summary, our algorithm produces a piecewise quadratic representation of f_K, which we can minimise efficiently to obtain $\hat{\theta}_K$. We also have piecewise linear representations of functions $b_2, \ldots, b_K$ through which we may iteratively obtain $\hat{\theta}_k = b_{k+1}(\hat{\theta}_{k+1})$ for $k = K-1, \ldots, 1$.

It seems challenging to obtain meaningful bounds on the number of computations that must be performed at each stage of this process in terms of parameters of the data. One can construct examples, albeit quite unrealistic, of f_k satisfying the properties in Lemma (2.4), such that f_{k+1} has more pieces than f_k by a factor. Following an inductive argument, this implies an exponential complexity, suggesting that a more useful result would require a suitable strengthening of the inductive hypothesis. However, to give an indication of the scalability of this algorithm, we ran a simple example with 3 true levels and found that with 50 categories the runtime was under 10^{-3} seconds; with 2000 categories it was still well under half a second. More details on computation time can be found in Sections 2.8.4 and 2.11.2. In Section 2.8.5, we describe an approximate version of the algorithm that can be used for fast computation in very large-scale settings.

2.3.2 Multivariate model

Using our dynamic programming algorithm for the univariate problem, we can attempt to minimise the objective (2.9) for the multivariate problem using block coordinate descent. This has been shown empirically to be a successful strategy for minimising objectives for high-dimensional regression with nonconvex penalties such as the MCP (Breheny and Huang, 2011, 2015; Mazumder et al., 2011), and we take this approach here. Considering the multivariate case, we iteratively minimise the objective Q over $\boldsymbol{\theta}_j = (\theta_{jk})_{k=1}^{K_j} \in \Theta_j$ keeping all other parameters fixed. Then for a given (γ, λ) and initial estimate $\hat{\boldsymbol{\theta}}^{(0)} \in \Theta$, we repeat the following until a suitable convergence criterion is met:

1. Initialise $m = 1$, and set for $i = 1, \ldots, n$

$$R_i = Y_i - \hat{\mu} - \sum_{l=1}^{p} \sum_{k=1}^{K_l} \hat{\theta}_{lk}^{(m-1)} \mathbb{1}_{\{X_{il}=k\}}.$$

2. For $j = 1, \ldots, p$, compute

$$R_i^{(j)} = R_i + \sum_{k=1}^{K_j} \hat{\theta}_{jk}^{(m-1)} \mathbb{1}_{\{X_{ij}=k\}} \qquad \text{for each } i, \tag{2.18}$$

$$\hat{\boldsymbol{\theta}}_j^{(m)} = \underset{\boldsymbol{\theta}_j \in \Theta_j}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(R_i^{(j)} - \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}_{\{X_{ij}=k\}} \right)^2 + \left(\sum_{k=1}^{K_j - 1} \rho_j(\theta_{j(k+1)} - \theta_{j(k)}) \right) \right\} \tag{2.19}$$

$$R_i = R_i^{(j)} - \sum_{k=1}^{K_j} \hat{\theta}_{jk}^{(m)} \mathbb{1}_{\{X_{ij}=k\}} \qquad \text{for each } i.$$

3. Increment $m \to m + 1$.

We define a blockwise optimum of Q to be any $\hat{\boldsymbol{\theta}} \in \Theta$, such that for each $j = 1, \ldots, p$,

$$\hat{\boldsymbol{\theta}}_j \in \underset{\boldsymbol{\theta}_j \in \Theta_j}{\arg\min}\, Q(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{j-1}, \boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_{j+1}, \ldots, \hat{\boldsymbol{\theta}}_p). \tag{2.20}$$

This is equivalent to $\hat{\boldsymbol{\theta}}$ being a fixed point of the block coordinate descent algorithm above. Provided $\gamma > 0$, Q is continuous in $\boldsymbol{\theta}$. As a consequence of Tseng (2001), Theorem 4.1 (c), provided the minimisers $\hat{\boldsymbol{\theta}}_j^{(m)}$ in (2.19) are unique for all j and m (which will invariably be the case when the responses are realisations of continuous random variables; see Proposition 2.3), then all limit points of the sequence $(\hat{\boldsymbol{\theta}}^{(m)})_{m=0}^{\infty}$ are blockwise optima. The application of this result is reliant on the global optimality of the solution returned by our dynamic programming algorithm for the univariate problem, and would not be possible were the univariate update step computed by any iterative optimisation procedure terminated in finite time.

2.3.3 Practicalities

In practice the block coordinate descent procedure described above must be performed over a grid of (γ, λ) values to facilitate tuning parameter selection by cross-validation. In line with analogous recommendations for other penalised regression optimisation procedures (Breheny and Huang, 2011; Friedman et al., 2010), we propose, for each fixed γ, to iteratively obtain solutions for an exponentially decreasing sequence of λ values, warm starting each application of block coordinate descent at the solution for the previous λ. It is our experience that this scheme speeds up convergence and helps to guide the resulting estimates to statistically favourable local optima by encouraging their sparsity, as has been shown theoretically for certain nonconvex settings (Wang et al., 2014).

The grid of γ values can be chosen to be fairly coarse as the solutions appear to be less sensitive to this tuning parameter; in fact fixing $\gamma \in \{8, 32\}$ yields competitive performance across a range of settings (see Section 2.6). The choice $\gamma \downarrow 0$, which mimics the ℓ_0 penalty,

has good statistical properties (see Theorem 2.5 and following discussion). However the global optimum typically has a smaller basin of attraction and can be prohibitively hard to locate, particularly in low signal to noise ratio settings where larger γ tends to dominate.

2.4 Theory

In this section, we study the theoretical properties of SCOPE. Recall our model

$$Y_i = \mu^0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk}^0 \mathbb{1}_{\{X_{ij}=k\}} + \varepsilon_i \tag{2.21}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\theta}^0 \in \Theta$. We will assume the errors $(\varepsilon_i)_{i=1}^n$ have mean zero, are independent and sub-Gaussian with parameter σ. Let

$$\Theta_0 = \left\{ \boldsymbol{\theta} \in \Theta : \theta_{jk} = \theta_{jl} \text{ whenever } \theta_{jk}^0 = \theta_{jl}^0 \text{ for all } j \right\}$$

and define the *oracle least squares estimate*

$$\hat{\boldsymbol{\theta}}^0 = \underset{\boldsymbol{\theta} \in \Theta_0}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \hat{\mu} - \sum_{j=1}^{p} \sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}_{\{X_{ij}=k\}} \right)^2 . \tag{2.22}$$

This is the least squares estimate of $\boldsymbol{\theta}^0$ with oracular knowledge of which categorical levels are fused in $\boldsymbol{\theta}^0$.

Note that in the case where the errors have equal variance v^2, the expected mean squared prediction error of $\hat{\boldsymbol{\theta}}^0$ satisfies

$$\mathbb{E}\left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\mu} - \mu^0 + \sum_{j=1}^{p} \sum_{k=1}^{K_j} (\hat{\theta}_{jk}^0 - \theta_{jk}^0) \mathbb{1}_{\{X_{ij}=k\}} \right)^2 \right\} \leq \frac{v^2}{n} \left(1 + \sum_{j=1}^{p} (s_j - 1) \right),$$

with equality when $\hat{\boldsymbol{\theta}}^0$ is unique. If the underlying model (while low-dimensional) is still reasonably large (e.g. $O(n)$), or if the error variance is large, there can be a meaningful difference between $\boldsymbol{\theta}^0$ and $\hat{\boldsymbol{\theta}}^0$. However, it is known from classical results that even with knowledge of which levels are fused in $\boldsymbol{\theta}^0$, the mean squared estimation error of $\hat{\boldsymbol{\theta}}^0$ cannot be improved upon within the class of linear unbiased estimators.

Our results below establish conditions under which $\hat{\boldsymbol{\theta}}^0$ is a blockwise optimum (2.20) of the SCOPE objective function Q (2.9), or in the univariate case when this in fact coincides with SCOPE. The minimum differences between the signals defined for each j by

$$\Delta(\boldsymbol{\theta}_j^0) = \min_{k,l} \left\{ |\theta_{jk}^0 - \theta_{jl}^0| : \theta_{jk}^0 \neq \theta_{jl}^0 \right\}, \tag{2.23}$$

will play a key role. If all components of $\boldsymbol{\theta}_j^0$ are equal we take $\Delta(\boldsymbol{\theta}_j^0)$ to be ∞. We also introduce $n_{j,\min} = \min_k n_{jk}$,

$$n_{j,\min}^0 = \min_k \sum_{l\,:\,\theta_{jl}^0 = \theta_{jk}^0} n_{jl} \qquad \text{and} \qquad n_{j,\max}^0 = \max_k \sum_{l\,:\,\theta_{jl}^0 = \theta_{jk}^0} n_{jl};$$

these latter two quantities are the minimum and maximum number of observations corresponding to a set of fused levels in the j^{th} predictor respectively.

2.4.1 Univariate model

We first consider the univariate case, where as usual we will drop the subscript j for simplicity. The following result establishes conditions for recovery of the oracle least squares estimate (2.22).

Theorem 2.5. *Consider the model* (2.21) *in the univariate case with* $p = 1$. *Suppose there exists* $\eta \in (0, 1]$ *such that* $\eta/s \leq n_{\min}^0/n \leq n_{\max}^0/n \leq 1/\eta s$. *Let* $\gamma_* = \min\{\gamma, \eta s\}$ *and* $\gamma^* = \max\{\gamma, \eta s\}$. *Suppose further that*

$$\Delta(\boldsymbol{\theta}^0) \geq 3\left(1 + \sqrt{2}/\eta\right)\sqrt{\gamma\gamma^*}\lambda. \tag{2.24}$$

Then with probability at least

$$1 - 2\exp\left(-\frac{n_{\min}\eta s\gamma_*\lambda^2}{8\sigma^2} + \log(K)\right), \tag{2.25}$$

the oracle least squares estimate $\hat{\boldsymbol{\theta}}^0$ (2.22) *is the global optimum of* (2.9), *so* $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$.

For a choice of the tuning parameters (γ, λ) with $\gamma \leq \eta s$ and λ such that equality holds in (2.24), we have, writing $\Delta \equiv \Delta(\boldsymbol{\theta}^0)$, that $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$ with probability at least

$$1 - 2\exp\left(-c\eta^2 n_{\min}\Delta^2/\sigma^2 + \log(K)\right),$$

where c is an absolute constant. The quantity η reflects how equal the number of observations in the true fused levels are: in settings where the prevalences of the underlying true levels are roughly equal, we would expect this to be closer to 1.

Consider now an asymptotic regime where K, s and $1/\Delta$ are allowed to diverge with n, $n_{\min} \asymp n/K$, so all levels have roughly the same prevalence, and η is bounded away from zero, so all true underlying levels also have roughly the same prevalence. Then in order for $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$ with high probability, we require $\Delta \gtrsim \sigma\sqrt{K\log(K)/n}$. This requirement cannot be weakened for any estimator; this fact comes as a consequence of minimax lower bounds on mis-clustering errors in Gaussian mixture models (Lu and Zhou, 2016, Theorem 3.3).

We remark that our result here concerning properties of the global minimiser of our objective is very different from existing results on local minimisers of objectives involving all-pairs-type penalties. For example, in the setting above where $K = n$, Theorem 2 of Ma and Huang (2017) gives that provided $s = o(n^{1/3}(\log n)^{-1/3})$ and $\Delta \gg \sigma s^{3/2} n^{-1/2}\sqrt{\log(n)}$, there exists a sequence of local minimisers converging to the oracle least squares estimate with high probability. This is significantly weaker than the condition $\Delta \gtrsim \sigma\sqrt{\log(n)}$ required for any estimator to recover oracle least squares in this setting, illustrating the substantial difference between results on local and global optima here.

2.4.2 Multivariate model

When the number of variables is $p > 1$, models can become high-dimensional, with ordinary least squares estimation failing to provide a unique solution. We will however assume that the solution for $\boldsymbol{\theta} \in \Theta_0$ to

$$\sum_{j=1}^{p}\sum_{k=1}^{K_j} \theta_{jk}^0 \mathbb{1}_{\{X_{ij}=k\}} = \sum_{j=1}^{p}\sum_{k=1}^{K_j} \theta_{jk} \mathbb{1}_{\{X_{ij}=k\}}$$

is unique, which occurs if and only if the oracle least squares estimate (2.22) is unique. In this case, we note that $\hat{\boldsymbol{\theta}}^0 = A\boldsymbol{Y}$ for a fixed matrix A. A necessary condition for this is that $\sum_j (s_j - 1) < n$.

Our result below provides a bound on the probability that the oracle least squares estimate is a blockwise optimum of the SCOPE objective (2.9) with $\rho_j = \rho_{\gamma_j,\lambda_j}$, recalling the definition of blockwise optima (2.20) in Section 2.3.2. This is much more meaningful than an equivalent bound for $\hat{\boldsymbol{\theta}}^0$ to be a local optimum as the number of local optima will be enormous. In general though there may be several blockwise optima, and it seems challenging to obtain a result giving conditions under which our blockwise coordinate descent procedure is guaranteed to converge to $\hat{\boldsymbol{\theta}}^0$. Our empirical results (Section 2.6) however show that the fixed points computed in practice tend to give good performance.

Theorem 2.6. *Consider the model* (2.21) *and assume* $\hat{\boldsymbol{\theta}}^0 = A\boldsymbol{Y}$. *Suppose that there exists* $\eta \in (0, 1]$ *such that* $\eta/s_j \leq n_{j,\min}^0/n \leq n_{j,\max}^0/n \leq 1/\eta s_j$ *for all* $j = 1, \ldots, p$. *Let* $\gamma_{*j} = \min\{\gamma_j, \eta s_j\}$ *and* $\gamma_j^* = \max\{\gamma_j, \eta s_j\}$. *Further suppose that*

$$\Delta(\boldsymbol{\theta}_j^0) \geq 3\left(\frac{4}{3} + \frac{\sqrt{2}}{\eta}\right)\sqrt{\gamma_j \gamma_j^*}\lambda_j. \tag{2.26}$$

Then letting $c_{\min} = (\max_l (AA^T)_{ll})^{-1}$, *with probability at least*

$$1 - 4\sum_{j=1}^{p} \exp\left(-\frac{(n_{j,\min} \wedge c_{\min})\eta\gamma_{*j} s_j \lambda_j^2}{8\sigma^2} + \log(K_j)\right), \tag{2.27}$$

the oracle least squares estimate $\hat{\boldsymbol{\theta}}^0$ *is a blockwise optimum of* (2.9).

Now suppose $\gamma_j \leq \eta s_j$ and λ_j are such that equality holds in (2.26) for all j. Then writing $K_{\max} = \max_j K_j$, $n_{\min} = \min_j n_{j,\min}$ and $\Delta_{\min} = \min_j \Delta(\boldsymbol{\theta}_j^0)$, we have that $\hat{\boldsymbol{\theta}}^0$ is a blockwise optimum of (2.9) with probability at least

$$1 - 4\exp\left(-c\eta^2(n_{\min} \wedge c_{\min})\Delta_{\min}^2/\sigma^2 + \log(K_{\max}p)\right),$$

where c is an absolute constant. Consider now an analogous asymptotic regime to that described in the previous section for the univariate case. Specifically assume $n_{\min} \asymp n/K_{\max}$ and $c_{\min} \gtrsim n_{\min}$ for simplicity. We then see that in order for $\hat{\boldsymbol{\theta}}^0$ to be a blockwise optimum with high probability, it is sufficient that $\Delta_{\min} \gtrsim \sigma\sqrt{K_{\max}\log(K_{\max}p)/n}$.

2.5 Extensions

In this section, we describe some extensions of our SCOPE methodology.

Continuous covariates. If some of the covariates are continuous rather than categorical, we can apply any penalty function of choice to these, and perform a regression by optimising the sum of a least squares objective, our SCOPE penalty and these additional penalty functions, using (block) coordinate descent.

For example, consider the model (2.1) with the addition of d continuous covariates. Let $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ be the centred design matrix for these covariates with i^{th} row $\boldsymbol{z}_i \in \mathbb{R}^d$. One can fit a model with SCOPE penalising the categorical covariates, and the Lasso with tuning parameter $\alpha > 0$ penalising the continuous covariates, resulting in the following objective over $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \Theta$:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu} - \boldsymbol{z}_i^T\boldsymbol{\beta} - \sum_{j=1}^{p}\sum_{k=1}^{K_j}\theta_{jk}\mathbb{1}_{\{X_{ij}=k\}}\right)^2 + \alpha\|\boldsymbol{\beta}\|_1 + \sum_{j=1}^{p}\sum_{k=1}^{K_j-1}\rho_j(\theta_{j(k+1)} - \theta_{j(k)}).$$

This sort of integration of continuous covariates is less straightforward when attempting to use tree-based methods to handle categorical covariates, for example.

Generalised linear models. Sometimes a generalised linear model may be appropriate. Although a quadratic loss function is critical for our exact optimisation algorithm described in Section 2.3.1, we can iterate local quadratic approximations to the loss term in the objective and minimise this. This results in a proximal Newton algorithm and is analogous to the standard approach for solving ℓ_1-penalised generalised linear models (Friedman et al., 2010, Section 3). An implementation of this scheme in the case of logistic regression for binary responses is available in the accompanying R package `CatReg`. We remark that when

computing logistic regression models with a SCOPE penalty it is advisable to use a larger value of γ than with a continuous response to aid convergence of the proximal Newton step; we recommend a default setting of $\gamma = 100$. In Section 2.6.2 we use the approach described above to perform a logistic regression using SCOPE on US census data.

Hierarchical categories. Often certain predictors may have levels that are effectively subdivisions of the levels of other predictors. Examples include category of item in e-commerce or geographical data with predictors for continent, countries and district. For simplicity, we will illustrate how such settings may be dealt with by considering a case with two predictors, but this may easily be generalised to more complex hierarchical structures. Suppose there is a partition $G_1 \cup \cdots \cup G_{K_1}$ of $\{1, \ldots, K_2\}$ such that for all $k = 1, \ldots, K_1$,

$$X_{i2} \in G_k \implies X_{i1} = k,$$

so the levels of the second predictor in G_k represent subdivisions of k^{th} level of the first predictor. Let $K_{2k} = |G_k|$ and let $\boldsymbol{\theta}_{2k}$ refer to the subvector $(\theta_{2l})_{l \in G_k}$ for each $k = 1, \ldots, K_1$, so components of $\boldsymbol{\theta}_{2k}$ are the coefficients corresponding to the levels in G_k. Also let $\theta_{2k(r)}$ denote the r^{th} order statistic within $\boldsymbol{\theta}_{2k}$. It is natural to encourage fusion among levels within G_k more strongly than for levels in different elements of the partition. To do this we can modify our objective function so the penalty takes the form

$$\sum_{k=1}^{K_1 - 1} \rho_1(\theta_{1(k+1)} - \theta_{1(k)}) + \sum_{k=1}^{K_1} \sum_{l=1}^{K_{2k}-1} \rho_{2k}(\theta_{2k(l+1)} - \theta_{2k(l)}).$$

We furthermore enforce the identifiability constraints that

$$\sum_{l=1}^{K_1} n_{1l}\theta_{1l} = 0 \quad \text{and} \quad \sum_{l \in G_k} n_{2l}\theta_{2l} = 0 \text{ for all } k = 1, \ldots, K.$$

As well as yielding the desired shrinkage properties, an additional advantage of this approach is that the least squares criterion is separable in $\boldsymbol{\theta}_{21}, \ldots, \boldsymbol{\theta}_{2K_1}$ so the blockwise update of $\boldsymbol{\theta}_2$ can be performed in parallel. This can lead to a substantial reduction in computation time if K_2 is large.

2.6 Numerical experiments

In this section we explore the empirical properties of SCOPE. We first present results on the performance on simulated data, and then in Sections 2.6.2 to 2.6.4 present analyses and experiments on US census data, insurance data and COVID-19 modelling data.

We denote SCOPE with a specific choice of γ as SCOPE-γ, and write SCOPE-CV to denote SCOPE with a cross-validated choice of γ. SCOPE solutions are computed using our R (R Core Team, 2019) package `CatReg` (Stokell et al., 2021a), using 5-fold cross-validation to select λ for all examples except those in Section 2.6.4. We compare SCOPE to linear or logistic regression where appropriate and a range of existing methods, including CAS-ANOVA (Bondell and Reich, 2009) (2.4), and an adaptive version where the weights $w_{j,kl}$ are multiplied by a factor proportional to $|\hat{\theta}^{\text{init}}_{jk} - \hat{\theta}^{\text{init}}_{jl}|^{-1}$, where $\hat{\boldsymbol{\theta}}^{\text{init}}$ is an initial CAS-ANOVA estimate. For these methods the tuning parameter λ was also selected by 5-fold cross-validation. As well as this, we include Delete or merge regressors (DMR) (Maj-Kańska et al., 2015) and Bayesian effect fusion (BEF) (Pauger and Wagner, 2019) in some experiments. With the former, models were fitted using `DMRnet` (Prochenka-Sołtys and Pokarowski, 2018) and selected by 5-fold cross-validation where possible; otherwise an information criterion was used. With BEF, coefficients were modelled with a Gaussian mixture model with posterior mean estimated using 1000 samples using `effectFusion` (Pauger et al., 2019). We also include comparison to the tree-based approaches CART (Breiman et al., 1984) and random forests (RF) (Breiman, 2001). Lastly, in some experiments, models were also fitted using the Lasso (Tibshirani, 1996). CART was implemented using `rpart` (Therneau and Atkinson, 2019) with pruning according to the one standard error rule. Random forests and Lasso were implemented using the default settings in `randomForest` (Liaw and Wiener, 2002) and `glmnet` (Friedman et al., 2010) packages respectively. For full details of the specific versions of these methods and software used in the numerical experiments, see Section 2.11.1.

2.6.1 Simulations

We simulated data according to the model (2.1) with the covariates X_{ij} generated randomly in the following way. We first drew $(W_{ij})_{j=1}^{p}$ from a multivariate $\mathcal{N}_p(0, \Sigma)$ distribution where the covariance matrix Σ had ones on the diagonal. The off-diagonal elements of Σ were chosen such that $U_{ij} = \Phi^{-1}(W_{ij})$ had $\text{corr}(U_{ij}, U_{ik}) = \rho$ for $j \neq k$. The marginally uniform U_{ij} were then quantised this to give $X_{ij} = \lceil 24U_{ij} \rceil$, so the number of levels $K_j = 24$.

The errors ε_i were independently distributed as $\mathcal{N}(0, \sigma^2)$. The performance of SCOPE and competitor methods was measured using mean squared prediction error on 10^5 new (noiseless) observations generated in the same way as the training data, and final results are averages over 500 draws of training and test data. We considered various settings of $(n, p, \rho, \boldsymbol{\theta}^0, \sigma^2)$ below with low-dimensional and high-dimensional scenarios considered in Sections 2.6.1 and 2.6.1 respectively. The coefficient vectors for each experiment are specified up to an additive constant, which is required to satisfy the identifiability condition (2.8).

We measured predictive performance by the mean squared prediction error (MSPE) given by

$$\text{MSPE} = \mathbb{E}_x\{g(x) - \hat{g}(x)\}^2, \tag{2.28}$$

where g is the true regression function, $\hat{g}$ an estimate, and the expectation is taken over the covariate vector x.

Low-dimensional experiments

Results are presented for three settings with $n = 500$, $p = 10$ given below.

1. $\boldsymbol{\theta}_j^0 = (\overbrace{-3,\ldots,-3}^{10\text{ times}}, \overbrace{0,\ldots,0}^{4\text{ times}}, \overbrace{3,\ldots,3}^{10\text{ times}})$ for $j = 1, 2, 3$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$.
2. $\boldsymbol{\theta}_j^0 = (\overbrace{-3,\ldots,-3}^{8\text{ times}}, \overbrace{0,\ldots,0}^{8\text{ times}}, \overbrace{3,\ldots,3}^{8\text{ times}})$ for $j = 1, 2, 3$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$.
3. As Setting 1, but with $\rho = 0.8$.

	Setting 1				Setting 2			
σ^2:	1	6.25	25	100	1	6.25	25	100
SNR:	4.7	1.9	0.95	0.47	4.2	1.7	0.85	0.42
SCOPE-8	**0.014**(0.0)	0.450(0.5)	4.571(1.0)	12.936(2.8)	**0.015**(0.0)	**0.285**(0.3)	6.775(0.9)	12.697(2.3)
SCOPE-32	0.018(0.0)	0.878(0.6)	4.151(0.9)	**12.356**(2.1)	0.019(0.0)	0.655(0.4)	5.026(1.0)	**12.037**(2.0)
SCOPE-CV	0.015(0.0)	**0.407**(0.4)	**4.120**(0.9)	12.513(2.5)	0.016(0.0)	0.292(0.3)	**5.005**(1.1)	12.444(2.5)
Linear regression	0.851(0.1)	5.317(0.7)	21.503(2.7)	86.745(10.7)	0.869(0.1)	5.406(0.7)	21.216(2.5)	85.439(10.9)
Oracle least squares	0.014(0.0)	0.091(0.1)	0.333(0.2)	1.405(0.8)	0.014(0.0)	0.088(0.0)	0.336(0.2)	1.532(0.8)
CAS-ANOVA	0.617(0.3)	1.602(0.3)	5.448(1.0)	14.814(2.2)	1.483(0.4)	1.626(0.3)	5.466(1.0)	13.421(2.2)
Adaptive CAS-ANOVA	0.135(0.1)	0.880(0.4)	5.076(1.2)	22.896(4.7)	0.134(0.1)	0.912(0.3)	5.535(1.2)	22.213(4.9)
DMR	**0.014**(0.0)	0.448(0.4)	4.884(1.4)	18.394(3.6)	0.016(0.0)	0.409(0.4)	6.430(1.4)	17.457(2.1)
BEF	0.020(0.0)	2.209(1.1)	6.297(1.8)	21.927(2.3)	0.019(0.0)	1.055(0.9)	8.183(2.0)	18.236(1.5)
CART	3.844(0.4)	5.099(0.9)	13.219(2.1)	22.431(1.2)	5.530(0.6)	7.457(0.9)	13.280(1.8)	18.198(0.7)
RF	9.621(0.5)	10.944(0.5)	13.217(0.7)	16.344(0.9)	8.947(0.3)	9.747(0.4)	11.249(0.6)	13.646(0.8)

	Setting 3			
σ^2:	1	6.25	25	100
SNR:	7.3	2.9	1.5	0.73
SCOPE-8	**0.015**(0.0)	0.967(0.7)	5.060(1.3)	14.555(2.9)
SCOPE-32	0.018(0.0)	0.713(0.4)	3.580(0.8)	**9.721**(1.9)
SCOPE-CV	0.022(0.1)	0.582(0.3)	**3.368**(0.9)	10.168(2.6)
Linear regression	0.879(0.1)	5.485(0.7)	21.987(2.7)	87.820(11.9)
Oracle least squares	0.014(0.0)	0.092(0.0)	0.362(0.2)	1.488(1.0)
CAS-ANOVA	0.710(0.2)	1.601(0.3)	4.732(0.9)	12.708(2.1)
Adaptive CAS-ANOVA	0.189(0.2)	0.701(0.3)	3.705(1.0)	16.186(3.6)
DMR	**0.015**(0.0)	**0.553**(0.5)	5.730(1.9)	18.594(4.5)
BEF	0.019(0.0)	1.716(0.9)	8.143(2.6)	26.923(7.0)
CART	4.336(0.6)	5.685(1.0)	9.910(1.7)	18.543(2.2)
RF	4.039(0.3)	5.673(0.5)	9.157(0.9)	13.766(1.7)

Table 2.1 Mean squared prediction errors (and standard deviations thereof) of various methods on the settings described.

Each of these experiments were performed with noise variance $\sigma^2 = 1$, 6.25, 25 and 100. Note that the variance of the signal varies across each setting, and signal-to-noise ratio (SNR) for each experiment is displayed in Table 2.1. Methods included for comparison were SCOPE-8, SCOPE-32, SCOPE-CV, linear regression, vanilla and adaptive CAS-ANOVA,

DMR, Bayesian effect fusion, CART and random forests. Also included are the results from the oracle least squares estimator (2.22).

Results are shown in Table 2.1 and further details are given in Section 2.11.2. Across all experiments, SCOPE with a cross-validated choice of γ exhibits prediction performance at least as good as the optimal approaches, and in all but the lowest noise settings performs better than the other methods that were included. In these exceptions, we see that fixing γ to be a small value (corresponding to high-concavity) provides leading performance.

In these low noise settings, we see that the methods based on first estimating the clusterings of the levels and then estimating the coefficients without introducing further shrinkage, such as DMR or Bayesian effect fusion, perform well. However they tend to struggle when the noise is larger. In contrast the tree-based methods perform poorly in low noise settings but exhibit competitive performance in high noise settings.

High-dimensional experiments

We considered 8 settings as detailed below, each with $n = 500$, $p = 100$ and simulated 500 times.

1. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{8\text{ times}},\overbrace{0,\ldots,0}^{8\text{ times}},\overbrace{2,\ldots,2}^{8\text{ times}})$ for $j = 1,2,3$, $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{10\text{ times}},\overbrace{0,\ldots,0}^{4\text{ times}},\overbrace{2,\ldots,2}^{10\text{ times}})$ for $j = 4,5,6$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 50$.

2. As Setting 1, but with $\rho = 0.5$.

3. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{8\text{ times}},\overbrace{0,\ldots,0}^{8\text{ times}},\overbrace{2,\ldots,2}^{8\text{ times}})$ for $j = 1,2,3$, $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{16\text{ times}},\overbrace{3,\ldots,3}^{8\text{ times}})$ for $j = 4,5,6$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0.5$ and $\sigma^2 = 100$.

4. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{5\text{ times}},\overbrace{-1,\ldots,-1}^{5\text{ times}},\overbrace{0,\ldots,0}^{4\text{ times}},\overbrace{1,\ldots,1}^{5\text{ times}},\overbrace{2,\ldots,2}^{5\text{ times}})$ for $j = 1,\ldots,5$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 25$.

5. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{16\text{ times}},\overbrace{3,\ldots,3}^{8\text{ times}})$ for $j = 1,\ldots,25$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 1$.

6. As Setting 5, but with $\rho = 0.5$.

7. $\boldsymbol{\theta}_j^0 = (\overbrace{-2,\ldots,-2}^{4\text{ times}},\overbrace{0,\ldots,0}^{12\text{ times}},\overbrace{2,\ldots,2}^{8\text{ times}})$ for $j = 1,\ldots,10$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 25$.

8. $\boldsymbol{\theta}_j^0 = (\overbrace{-3,\ldots,-3}^{6\text{ times}},\overbrace{-1,\ldots,-1}^{6\text{ times}},\overbrace{1,\ldots,1}^{6\text{ times}},\overbrace{3,\ldots,3}^{6\text{ times}})$ for $j = 1,\ldots,5$, and $\boldsymbol{\theta}_j^0 = 0$ otherwise; $\rho = 0$ and $\sigma^2 = 25$.

Models were fitted using SCOPE-8, SCOPE-32, SCOPE-CV, DMR, CART, Random forests and the Lasso. Table 2.2 gives the mean squared prediction errors across each of the settings.

As well as prediction performance, it is interesting to see how the methods perform in terms of variable selection performance. With categorical covariates, there are two potential ways of evaluating this. The first is to consider the number of false positives and false negatives across the $p = 100$ categorical variables, defining a variable j to have been selected if $\hat{\boldsymbol{\theta}}_j \neq 0$. These results are shown in Table 2.3. This definition of a false positive can be considered quite conservative; typically one can find that often the false signal variables have only two levels, each with quite small coefficients. This means that the false positive rate can increase substantially with only a small increase in the dimension of the estimated linear model.

The second is to see within the signal variables (i.e., the j for which $\boldsymbol{\theta}^0_j \neq 0$), how closely the estimated clustering resembles the true structure. To quantify this, we use the *adjusted Rand index* (Hubert and Arabie, 1985). This is the proportion of all pairs of observations that are either (i) in different true clusters and different estimated clusters, or (ii) in the same true cluster and estimated cluster; this is then corrected to ensure that its value is zero when exactly one of the clusterings is 'all-in-one'. In Table 2.4 we report the average adjusted Rand index over the true signal variables in each setting.

Setting:	1	2	3	4	5	6	7	8
SNR:	0.6	1.0	1.0	0.64	12	36	0.87	1.0
SCOPE-8	14.319(2.0)	15.445(2.9)	30.597(5.6)	7.254(1.2)	96.538(25.0)	7.960(23.2)	15.867(1.4)	11.028(1.6)
SCOPE-32	**14.009**(1.6)	**10.780**(1.6)	**21.841**(3.4)	7.256(0.9)	65.344(13.4)	0.107(0.0)	14.867(1.2)	11.218(1.4)
SCOPE-CV	14.026(1.7)	10.843(1.8)	22.004(3.9)	**7.191**(1.0)	**54.030**(19.2)	**0.084**(0.0)	**14.865**(1.3)	**10.941**(1.5)
Oracle LSE	5.044(0.6)	5.130(0.6)	2.664(1.0)	1.09 (0.3)	0.054(0.0)	0.055(0.0)	1.087(0.3)	0.799(0.3)
DMR	18.199(1.4)	22.627(4.4)	42.979(9.2)	9.645(1.2)	139.095(4.3)	213.691(35.7)	19.298(0.8)	11.737(2.4)
CART	18.146(0.5)	31.235(3.6)	58.73 (6.6)	10.466(0.3)	139.35 (2.1)	614.739(42.8)	19.021(0.4)	23.775(1.5)
RF	16.181(0.6)	16.345(1.4)	31.561(2.6)	9.053(0.4)	128.618(2.2)	264.374(14.4)	17.224(0.4)	19.783(0.7)
Lasso	18.136(0.5)	24.839(1.3)	48.162(2.5)	10.473(0.4)	135.375(5.0)	154.656(7.8)	18.886(0.6)	23.813(1.6)

Table 2.2 Mean squared prediction errors (and standard deviations thereof) of each of the methods in the 8 high-dimensional settings considered.

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	0.02/0.35	0.04/0.23	0.04/0.25	0.02/0.15	0.02/0.23	0.02/0.01	0.02/0.35	0.01/0.00
SCOPE-32	0.14/0.15	0.30/0.02	0.30/0.02	0.15/0.04	0.52/0.00	0.00/0.00	0.21/0.08	0.21/0.00
SCOPE-CV	0.12/0.20	0.30/0.02	0.29/0.03	0.12/0.07	0.59/0.00	0.00/0.00	0.21/0.11	0.09/0.00
DMR	0.00/0.86	0.00/0.44	0.00/0.47	0.00/0.62	0.00/0.91	0.03/0.60	0.00/0.88	0.00/0.02
Lasso	0.01/0.88	0.00/1.00	0.00/1.00	0.01/0.83	0.00/0.98	0.00/1.00	0.00/0.91	0.00/0.90

Table 2.3 (False positive rate)/(False negative rate) of linear modelling methods considered in the high-dimensional settings.

Further details can be found in Section 2.11.2. In particular we include a table with the distribution of cross-validated choices of γ (from a grid $\{4, 8, 16, 32, 64\}$) for each experimental

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	0.23	0.36	0.38	0.15	0.39	0.96	0.13	0.29
SCOPE-32	0.29	0.46	0.48	0.19	0.56	1.00	0.17	0.34
SCOPE-CV	0.27	0.45	0.46	0.18	0.56	1.00	0.17	0.31
DMR	0.04	0.20	0.23	0.06	0.04	0.19	0.03	0.28
Lasso	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2.4 Average adjusted Rand index among true signal variables for the high-dimensional settings.

setting. Note that a choice of $\gamma = 4$ is close to the setting of $\gamma = 3$ recommended in Zhang (2010), though the problem of categorical covariates is very different in nature than the vanilla variable selection problem considered there. Our results there suggest that for SCOPE, a larger value of γ is preferable across a range of settings, which is also visible in the comparison between $\gamma = 8$ and $\gamma = 32$ in Table 2.2.

Across all the settings in this study, SCOPE performs better than any of the other methods included. This is regardless of which of the three γ regimes is chosen, although cross-validating γ gives the strongest performance overall. Comparing the results for $\gamma = 8$ and $\gamma = 32$ suggests that a larger (low-concavity) choice of γ is preferable for higher-dimensional settings. In setting 6, we see from Tables 2.3 and 2.4 that SCOPE obtains the true underlying groupings of the coefficients and obtains the oracle least squares estimate in every case, giving these striking results. This is also achieved for some of the experiments in setting 5. In contrast, DMR, which initially applies a group Lasso (Yuan and Lin, 2006) to screen the categorical variables and give a low-dimensional model, necessarily misses some signal variables in this first stage and hence struggles here.

2.6.2 Adult dataset examples

The *Adult dataset*, available from the UCI Machine Learning Repository (Dua and Graff, 2019), contains a sample of 45 222 observations based on information from the 1994 US census. The binary response variable is 0 if the individual earns at most \$50 000 a year, and 1 otherwise. There are 2 continuous and 8 categorical variables; some such as 'native country' have large numbers of levels, bringing the total dimension to 93. An advantage of using SCOPE here over black-box predictive tools such as Random forests is the interpretability of the fitted model.

In Table 2.5, we show the 25-dimensional fitted model. Within the Education category, we see that six distinct levels have been identified. These agree almost exactly with the stratification one would expect, with all school dropouts before 12th grade being grouped together at the lowest level.

Here we assess performance in the challenging setting when the training set is quite small by randomly selecting 1% (452) of the total observations for training, and using the

Variable	Coefficient	Levels
Intercept	−3.048	–
Age	0.027	–
Hours per week	0.029	–
Work class	0.378	Federal government, Self-employed (incorporated)
	0.058	Private
	−0.143	Local government
	−0.434	Self-employed (not incorporated), State government, Without pay
Education level	1.691	Doctorate, Professional school
	1.023	Master's
	0.646	Bachelor's
	−0.132	Associate's (academic), Associate's (vocational), Some college (non-graduate)
	−0.546	12th, High school grad
	−1.539	Preschool, 1st-4th, 5th-6th, 7th-8th, 9th, 10th, 11th
Marital status	0.059	Divorced, Married (armed forces spouse), Married (civilian spouse), Married (absent spouse), Separated, Widowed
	−0.476	Never married
Occupation	0.560	Executive/Managerial
	0.311	Professional/Specialty, Protective service, Tech support
	−0.003	Armed forces, Sales
	−0.168	Admin/Clerical, Craft/Repair
	−0.443	Machine operative/inspector, Transport
	−1.107	Farming/Fishing, Handler/Cleaner, Other service, Private house servant
Relationship*	1.498	Wife
	0.332	Husband
	−1.220	Not in family
	−1.482	Unmarried, Other relative
	−2.144	Own child
Race	0.013	White
	0.008	Asian/Pacific islander, Other
	−0.182	Native-American/Inuit, Black
Sex	0.139	Male
	−0.619	Female
Native country	0.018	KH, CA, CU, ENG, FR, DE, GR, HT, HN, HK, HU, IN, IR, IE, IT, JM, JP, PH, PL, PT, PR, TW, US, YU
	−0.882	CN, CO, DO, EC, SV, GT, NL, LA, MX, NI, GU-VI-etc, PE, SCT, ZA, TH, TT, VN

Table 2.5 Coefficients of SCOPE model trained on the full dataset. Here, $\gamma = 100$ and λ was selected by 5-fold cross-validation (with cross-validation error of 16.82%). Countries, aside from those in the UK, are referred to by their (possibly historical) internet top-level domains. *Relation with which the subject lives.

remainder as a test set. Any observations containing levels not in the training set were removed. Models were fitted with SCOPE-100, SCOPE-250, logistic regression, vanilla and adaptive CAS-ANOVA, DMR, Bayesian effect fusion, CART and random forests.

We see that both SCOPE-100 and SCOPE-250 are competitive, with CART and Random forests also performing well, though the latter two include interactions in their fits. CAS-ANOVA also performs fairly well, the misclassification error is larger that for both versions of SCOPE, and the average fitted model size is larger.

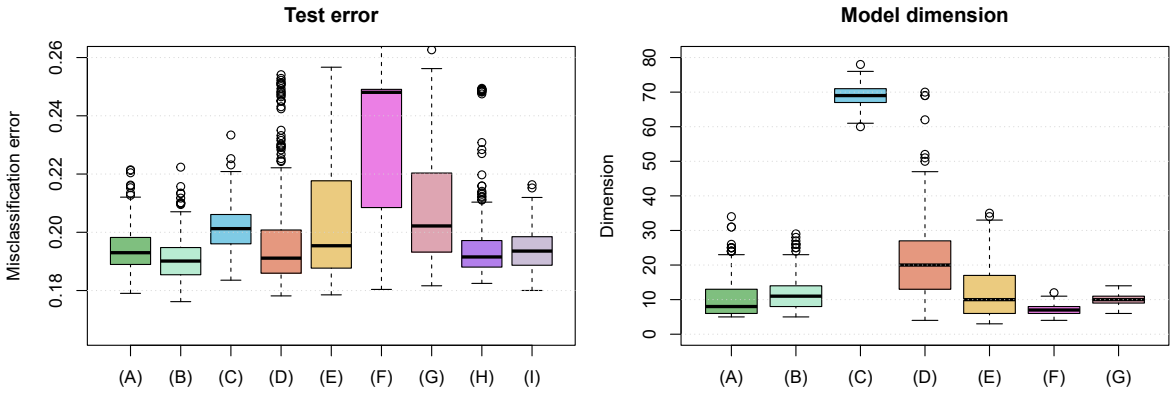


Figure 2.4 Prediction performance and fitted model dimension (respectively) of various methods on the Adult dataset: (A) SCOPE-100; (B) SCOPE-250; (C) Logistic regression; (D) CAS-ANOVA; (E) Adaptive CAS-ANOVA; (F) DMR; (G) BEF; (H) CART; (I) RF.

Method	Misclassification error	Model dimension	Computation time (s)
SCOPE-100	0.194	10.5	467
SCOPE-250	**0.191**	11.8	450
Logistic regression	0.202	68.9	0.04
CAS-ANOVA	0.198	21.5	429
Adaptive CAS-ANOVA	0.205	11.7	8757
DMR	0.235	6.9	11
BEF	0.207	9.8	1713
CART	0.196		0.01
RF	0.194		0.14

Table 2.6 Results of experiments on the Adult dataset.

Adult dataset with artificially split levels

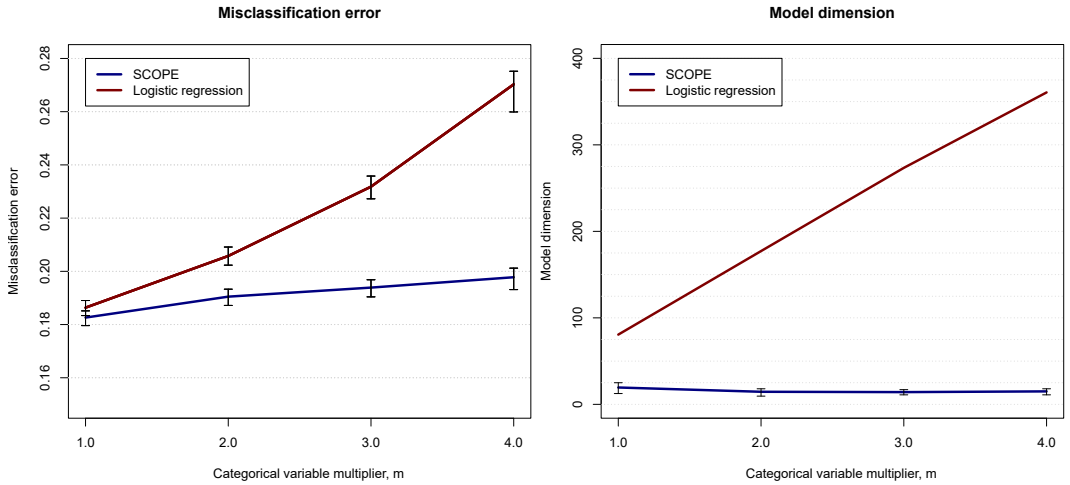


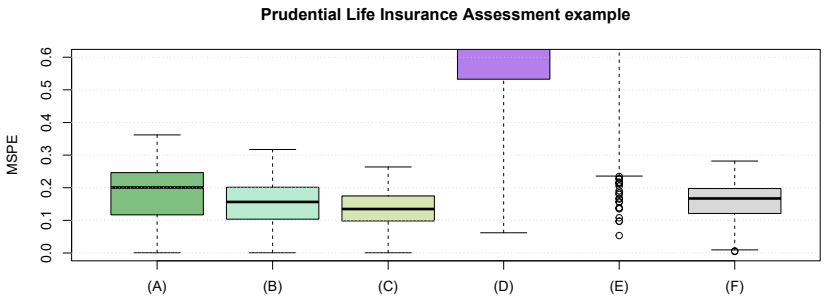
Figure 2.5 Misclassification error and dimensions of models fitted on a sample of the *Adult dataset* when levels have been artificially split m times.

To create a more challenging example, we artificially created additional levels in the *Adult dataset* as follows. For each categorical variable we recursively selected a level with probability proportional to its prevalence in the data and then split it into two by appending "-0" or "-1" to the level for each observation independently and with equal probabilities. We repeated this until the total number of levels reached m times the original number of levels for that variable for $m = 2, 3, 4$. This process simulates for example responses to a survey, where different respondents might answer 'US', 'U.S.', 'USA', 'U.S.A.', 'United States' or 'United States of America' to a question, which would naively all be treated as different answers.

We used 2.5% (1130) of the observations for training and the remainder for testing and applied SCOPE with $\gamma = 100$ and logistic regression. Results were averaged over 250 training and test splits. Figure 2.5 shows that as the number of levels increases, the misclassification error of SCOPE increases only slightly and the fitted model dimension remains almost unchanged, whereas both increase with m for logistic regression.

2.6.3 Insurance data example

The Prudential Life Insurance Assessment challenge was a prediction competition run on Kaggle. By more accurately predicting risk, the burden of extensive tests and check-ups for life insurance policyholders could potentially be reduced. For this experiment, we use the training set that was provided for entrants of the competition.



Method	MSPE
SCOPE-8	0.265
SCOPE-32	0.244
SCOPE-CV	**0.211**
CART	0.851
RF	1.120
Lasso	0.244

Figure 2.6 Mean squared prediction error on the example based on the Prudential Life Insurance Assessment dataset. Methods used are: (A) SCOPE-8; (B) SCOPE-32; (C) SCOPE-CV; (D) CART; (E) RF; (F) Lasso.

We removed a small number of variables due to excessive missingness, leaving 5 continuous variables and 108 categorical variables, most with 2 or 3 levels but with some in the hundreds (and the largest with 579 levels). Rather than using the response from the original dataset, which is ordinal, to better suit the regression setting we are primarily concerned with in this work, we artificially generated a continuous response. To construct this signal, firstly 10 of the categorical variables were selected at random, with probability proportional to the number of levels. For the j^{th} of these, writing K_j for the number of levels, we set $s_j = \lfloor 2 + \frac{1}{2} \log K_j \rfloor$ and assigned each level a coefficient in $1, \ldots, s_j$ uniformly at random, thus yielding s_j true

levels. The coefficients for the 5 continuous covariates were generated as draws from $\mathcal{N}_5(0, I_5)$. The response was then scaled to have unit variance, after which standard normal noise was added.

We used 10% ($n = 5938$) of the 59 381 total number of observations for training, and the remainder to compute an estimated MSPE (2.28) by taking an average over these observations. We repeated this 1000 times, sampling 10% of the observations and generating the coefficients as above anew in each repetition. The average mean squared prediction errors achieved by the various methods under comparison are given in Figure 2.6. We see that SCOPE with a cross-validated choice of γ performs best, followed by the Lasso and SCOPE-32.

2.6.4 COVID-19 Forecast Hub example

As well as the prediction performance experiments in the rest of this section, we include an exploratory data analysis example based on data relating to the ongoing (at time of writing) global COVID-19 pandemic. The COVID-19 Forecast Hub (2020) *'. . . serves as a central repository of forecasts and predictions from over 50 international research groups.'* A collection of different research groups publish forecasts every week of case incidence in each US state for some number of weeks into the future.

In order to understand some of the difficulties of this challenging forecasting problem, we fitted an error decomposition model of the form

$$\log\left(\frac{1 + \text{cases}_{w,l}}{1 + \text{est.cases}_{m,t,w,l}}\right) = \alpha_0 + \alpha_{m,t} + \beta_{w,l} + \eta_{m,t,w,l}, \tag{2.29}$$

where w is the week that the forecast is for, l is the state, m indexes the forecasting model, t is the 'target' number of weeks in the future the forecast is for, $\eta_{m,t,w,l}$ is an error term, and $\text{cases}_{w,l}$ and $\text{est.cases}_{m,t,w,l}$ are the observed and estimated cases respectively. This decomposition allows an interaction term between time and location, which is important given that the pandemic was known to be more severe at different times for different areas. An interaction between model and forecasting distance was also included in order to capture the effect of some models potentially being more 'short-sighted' than others. The inclusion of the $+1$ on the left-hand side is to avoid numerators or denominators of zero.

We used data from 6 April 2020 to 19 October 2020, giving a total of 100 264 (m, t, w, l)-tuples. We applied a SCOPE penalty with $\gamma = 8$ to $\beta_{w,l}$, which had 1428 levels. The $\alpha_{m,t}$ coefficients, which amounted to 170 levels, were left unpenalised. The additional tuning parameter λ was selected using the Extended Bayesian Information Criterion (Chen and Chen, 2008) rather than cross-validation, as it was more suited to this sort of exploratory analysis on data with a chronological structure.

The resulting estimates $\hat{\beta}_{w,l}$ had 8 levels. We measured the 'similarity' of two US states l_a and l_b over a period of time by computing the proportion of weeks at which their estimates

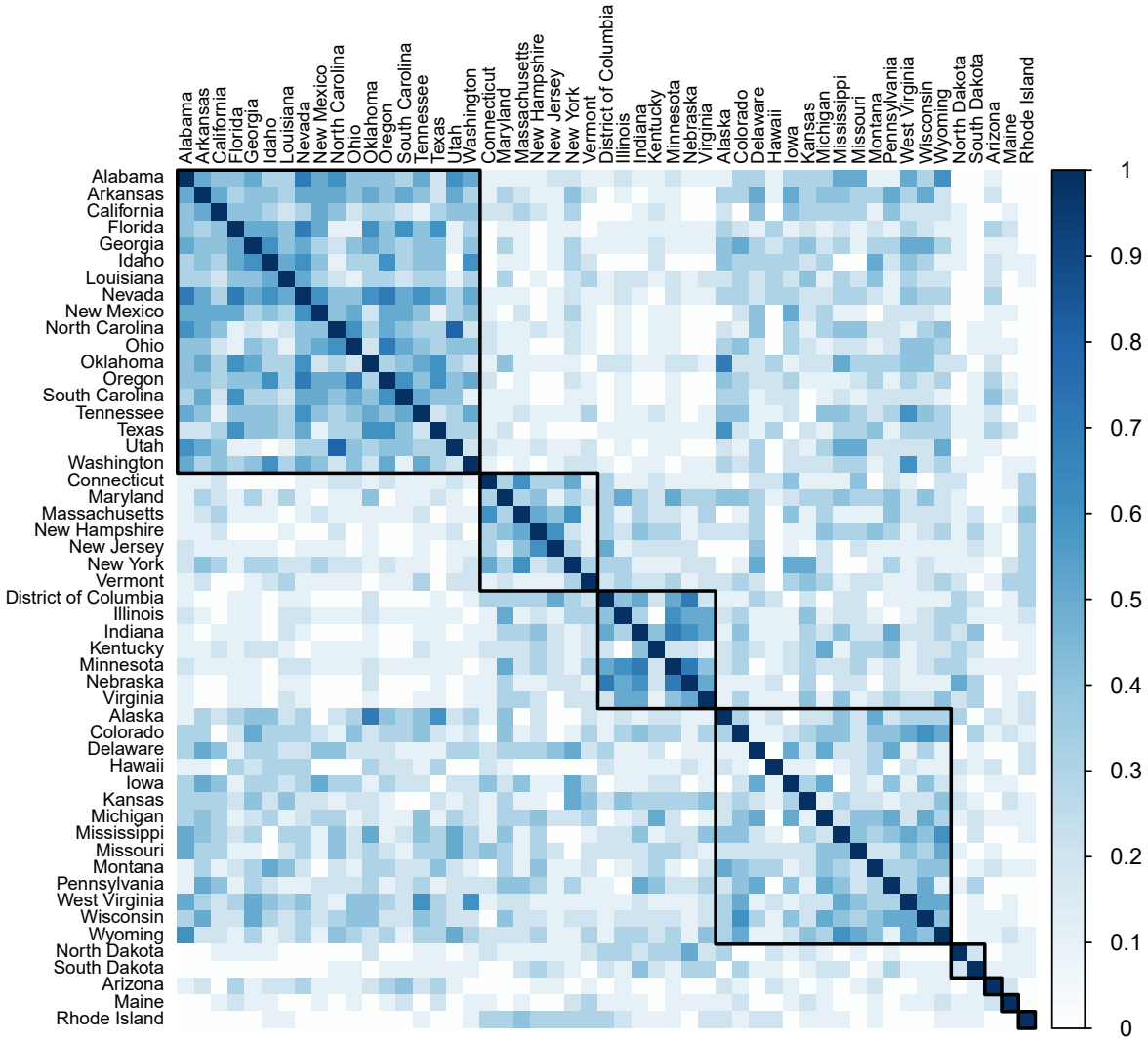


Figure 2.7 Similarity matrix for US states computed based on data relating to the second 'wave' of the COVID-19 pandemic in the US, taken to be from 26 June 2020 to 29 August 2020.

$\hat{\beta}_{w,l_a} = \hat{\beta}_{w,l_b}$ coincided. The similarity matrix presented in Figure 2.7 was constructed based on the second 'wave' of the epidemic which occurred in Summer 2020, with clusters identified by applying spectral clustering on the similarity matrix and plotted in order of decreasing within-cluster median pairwise similarity.

The resulting clusters are at once interpretable and interesting. Roughly speaking, the top 3 clusters ('top' when ordered according to median pairwise within-cluster agreement) correspond to states that experienced notable pandemic activity in the second, first, and third 'waves' of the U.S. coronavirus pandemic, respectively. The first cluster features several southern States (e.g., Georgia, Florida, Texas) which experienced a surge of COVID cases in June–July. The second cluster features east coast states (e.g., New Jersey and New

York) which experienced an enormous pandemic toll in April–May. And the third features midwestern states (e.g., Kentucky, Indiana, Nebraska) which had upticks most recently in September-October.

2.7 Discussion

In this work we have introduced a new penalty-based method for performing regression on categorical data. An attractive feature of a penalty-based approach is that it can be integrated easily with existing methods for regression with continuous data, such as the Lasso. Our penalty function is nonconvex, but in contrast to the use of nonconvex penalties in standard high-dimensional regression problems, the nonconvexity here is necessary in order to obtain sparse solutions, that is fusions of levels. Whilst computing the global optimum of nonconvex problems is typically very challenging, for the case with a single categorical variable with several hundred levels, our dynamic programming algorithm can typically solve the resulting optimisation problem in less than a second on a standard laptop computer. The algorithm is thus fast enough to be embedded within a block coordinate descent procedure for handling multiple categorical variables.

We give sufficient conditions for SCOPE to recover the oracle least squares solution when $p = 1$ involving a minimal separation between unequal coefficients that is optimal up to constant factors. For the multivariate case where $p > 1$, we show that oracle least squares is a fixed point of our block coordinate descent algorithm, with high probability.

Our work offers several avenues for further work. On the theoretical front, it would be interesting to obtain guarantees for block coordinate descent to converge to a local optimum with good statistical properties, a phenomenon that we observe empirically. On the methodology side, it would be useful to generalise the penalty to allow for clustering multivariate coefficient vectors; such clustering could be helpful in the context of mixtures of regressions models, for example.

2.8 Supplementary material for Section 2.3

2.8.1 Candidate minimiser functions

In this section we give explicit forms of the functions $p_{k,r}$ as defined in Section 2.3.1. We write $q_{k,r}(x) = a_r x^2 + b_r x + c_r$ for simplicity, suppressing the subscript k. For $S \subseteq \mathbb{R}$ and $a, b \in \mathbb{R}$, we write $aS + b$ for the set $\{ax + b : x \in S\}$.

Recall from Section 2.3.1 that

$$u_{k,r,t}(\theta_{k+1}) = \tilde{\min}_{\theta_k \in D_k : \theta_k < \theta_{k+1}} \{\tilde{q}_{k,r}(\theta_k) + \tilde{\rho}_t(\theta_{k+1} - \theta_k)\}.$$

For a function $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$, we denote the *effective domain* of f by

$$\text{dom } f = \{x \in \mathbb{R} : f(x) < \infty\}.$$

For each $r = 1, \ldots, m(k)$, there are cases corresponding to $t = 1$ and $t = 2$. The formulas are as follows:

$$u_{k,r,1}(x) = \frac{2a_r x^2 + 2(b_r - 2a_r\gamma\lambda)x + (b_r - 2a_r\gamma\lambda)^2}{2(1 - 2a_r\gamma)} + c_r,$$

$$\text{with} \quad \text{dom } u_{k,r,1} = \begin{cases} ((1 - 2a_r\gamma)I_{k,r} + \gamma(\lambda - b_r)) \cap \left[\frac{4a_r\gamma\lambda - b_r}{2a_r}, \frac{\lambda - b_r}{2a_r}\right) & \text{if} \quad 2a_r - 1/\gamma > 0 \\ \emptyset & \text{otherwise.} \end{cases}$$

If $g_k(\theta_{k+1}) = u_{k,r,1}(\theta_{k+1})$, then

$$b_k(\theta_{k+1}) = \frac{\theta_{k+1} + \gamma(b_r - \lambda)}{1 - 2a_r\gamma}.$$

The second case is

$$u_{k,r,2}(x) = -\frac{b_r^2}{4a_r} + c + \frac{1}{2}\gamma\lambda^2,$$

$$\text{with} \quad \text{dom } u_{k,r,2} = \begin{cases} \left[-\frac{b_r}{2a_r} + \gamma\lambda, \infty\right) & \text{if} \quad a_r > 0 \text{ and } -b_r/2a_r \in I_{k,r} \\ \emptyset & \text{otherwise.} \end{cases}$$

Here, if $g_k(\theta_{k+1}) = u_{k,r,2}(\theta_{k+1})$, then

$$b_k(\theta_{k+1}) = -b_r/2a_r.$$

Considering (2.16), we see that we can also have the case where $g_k(\theta_{k+1}) = f_k(\theta_{k+1})$. Thus we can form the set of quadratics $p_{k,r}$ and associated intervals as the set of $u_{k,r,t}$ as above for $t = 1, 2$ and the $q_{k,r}$ themselves. Note that when $g_k(\theta_{k+1}) = q_{k,r}(\theta_{k+1})$, we have $b_k(\theta_{k+1}) = \theta_{k+1}$.

2.8.2 Algorithm details

Algorithm 2.1 describes in detail how the optimisation routine works. In the algorithm we make use of the following objects:

- for $x \in \mathbb{R}$, $A(x)$ is the active set at x;
- E is the set of points at which the active set changes;
- $N(x)$ is the intersection set at x;

Algorithm 2.1 Outline of procedure for computing f_k

while $E, N(x) \neq \emptyset$ **do**
  if $\min\{y\colon (y,r) \in N(x)\} < \min E$ **then**
    $(y^*, r^*) = \arg\min\{y\colon (y,r) \in N(x)\}$
    $U = U \cup \{([\tilde{x}, y^*), r(x))\}$, $x = \tilde{x} = y^*$, $r(x) = r^*$
    $N(x) = \emptyset$, for any intersection between $p_{k-1,r(x)}$ and any $p_{k-1,r}$ with $r \in A(x)\backslash\{r(x)\}$ at location $y > x$, set $N(x) = N(x) \cup \{(y,r)\}$.
  else
    $y^* = \min E$, $E = E\backslash\{y^*\}$,
    Update active set $A(y^*)$
    if $r(x) \notin A(y^*)$ **then**
      Set r^* such that $p_{k-1,r^*} = \mathsf{ChooseFunction}(A(y^*), y^*)$
      $U = U \cup \{([\tilde{x}, y^*), r(x))\}$, $x = \tilde{x} = y^*$, $r(x) = r^*$
      $N(x) = \emptyset$, for any intersection between $p_{k-1,r(x)}$ and any $p_{k-1,r}$ with $r \in A(x)\backslash\{r(x)\}$ at location $y \geq x$, set $N(x) = N(x) \cup \{(y,r)\}$.
    else
      if $p_{k-1,r(x)} \neq p_{k-1,r^*} = \mathsf{ChooseFunction}(A(y^*), y^*)$ **then**
        $U = U \cup \{([\tilde{x}, y^*), r(x))\}$, $x = \tilde{x} = y^*$, $r(x) = r^*$
        $N(x) = \emptyset$, for any intersection between $p_{k-1,r(x)}$ and any $p_{k-1,r}$ with $r \in A(x)\backslash\{r(x)\}$ at location $y > x$, set $N(x) = N(x) \cup \{(y,r)\}$.
      else
        if $A(y^*) \neq A(x)$ **then**
          For any intersection between $r(x)$ and any $r \in A(y^*)\backslash A(x)$ at location $y > x$, set $N(y^*) = N(y^*) \cup \{(y,r)\}$.
          For any $(y,r) \in N(x)$ with $r \notin A(y^*)$, set $N(y^*) = N(y^*)\backslash\{(y,r)\}$
          $x = y^*$
        end if
      end if
    end if
  end if
end while

Algorithm 2.2 ChooseFunction(H, x)

Input: $H = \{h_1, \ldots, h_n\}$ a set of functions, x a real number
Set $H_1 = \arg\min\{h(x) : h \in H\}$
if $|H_1| = 1$ **then**
  Select $h^* \in H_1$
else
  Set $H_2 = \arg\min\{h'(x) : h \in H_1\}$
  if $|H_2| = 1$ **then**
    Select $h^* \in H_2$
  else
    Set $H_3 = \arg\min\{h''(x) : h \in H_2\}$
    Select $h^* \in H_3$ (choosing $h_i \in H_3$ with i minimal if $|H_3| > 1$)
  end if
end if
Output: h^*

- U is a set of tuples (I, r) where $I \subseteq \mathbb{R}$ is an interval and r is an integer, which is dynamically updated as the algorithm progresses.

See Section 2.3.1 for definitions of the sets above. We also use the convention that if $x = -\infty$ then $[x, y) = (-\infty, y)$.

All of the $p_{k,1}, \ldots, p_{k,m(k)}$ and $J_{k,m}$ are computed at the start of each iterate k. We then initialise

$$E = \bigcup_{r=1}^{n(k)} \partial J_{k-1,r},$$

the set of all of the end-points of the intervals $J_{k-1,1}, \ldots, J_{k-1,n(k)}$.

Here x can be thought of as the 'current position' of the algorithm; $\tilde{x}$ is used to store when the minimising function $p_{k-1,r(x)}$ last changed. We initialise $\tilde{x} = -\infty$ and $x = -1 + \max\{y \in I_{k-1,1} \colon f'_{k-1}(y_-) \leq 0\}$. This choice of x ensures that the active set $A(x)$ contains only one element (as mentioned in Section 2.3.1); this will always be the index corresponding to the function $\tilde{q}_{k-1,1}$.

We initialise the output set $U = \emptyset$, which by the end of this algorithm will be populated with the functions $\tilde{q}_{k,1}, \ldots, \tilde{q}_{k,m(k)}$ and their corresponding intervals $I_{k,1}, \ldots, I_{k,m(k)}$ that partition $\mathbb{R}$. Finally, we initialise the set $N(x)$ which will contain the intersections between $p_{k-1,r(x)}$ and other functions in the active set. As the active set begins with only one function, we set $N(x) = \emptyset$.

As mentioned in Section 2.3.1, there are several modifications that can speed up the algorithm. One such modification follows from the fact that for each r, $u_{k,r,2}$ is a constant function over its effective domain, and their effective domain is a semi-infinite interval (see

Section 2.8.1 for their expressions). Therefore, for a given point $x \in \mathbb{R}$, we can remove all such functions from $A(x)$ except for the one taking the minimal value.

We also note that in Algorithm 2.1, the set $N(x)$ is not recomputed in its entirety at every point x at which $A(x)$ is updated, as is described in Section 2.3.1. Line 13 shows how sometimes $N(x)$ can instead be updated by adding or removing elements from it. Often, points 3 (i) and 3 (ii) from the description in the Section 2.3.1 will coincide, and in such instances some calls to ChooseFunction (Algorithm 2.2) can be skipped.

2.8.3 Remarks on formulations of the univariate objective

It is clear why the identifiability constraint (2.8) is important when we consider the multivariate problem in Section 2.3.2. However, for the univariate problem, both constrained and unconstrained formulations of the objective can be clearly defined:

$$\hat{\boldsymbol{\theta}}^{\mathrm{c}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{1}{2} \sum_{k=1}^{K} w_k \left(\bar{Y}_k - \hat{\mu} - \theta_k \right)^2 + \sum_{k=1}^{K-1} \rho(\theta_{(k+1)} - \theta_{(k)}) \right\}, \tag{2.30}$$

$$\hat{\boldsymbol{\theta}}^{\mathrm{u}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^K} \left\{ \frac{1}{2} \sum_{k=1}^{K} w_k \left(\bar{Y}_k - \theta_k \right)^2 + \sum_{k=1}^{K-1} \rho(\theta_{(k+1)} - \theta_{(k)}) \right\}. \tag{2.31}$$

As discussed in Section 2.3.1, we can enlarge the feasible set in (2.30) to be all of $\mathbb{R}^K$: similarly to the observation that $\sum_k w_k \hat{\theta}^{\mathrm{u}}_k = \hat{\mu} = \sum_k w_k \bar{Y}_k$, the minimiser of (2.30) over all of $\mathbb{R}^K$ will always be in Θ. This can be shown by following the argument at the beginning of the proof of Lemma 2.10. Therefore the algorithm defined in Section 2.3.1 can also be applied to the unconstrained formulation of the objective.

It is clear that these problems are essentially identical, as $\hat{\boldsymbol{\theta}}^{\mathrm{u}}$ is a minimiser of the unconstrained objective if and only if $\hat{\boldsymbol{\theta}}^{\mathrm{u}} - \hat{\mu}\mathbf{1}$ is a minimiser of the constrained objective. Observe that while $\hat{\boldsymbol{\theta}}^{\mathrm{u}} \in \mathbb{R}^K$, the solution to the constrained objective is in fact $(\hat{\mu}, \hat{\boldsymbol{\theta}}^{\mathrm{c}}) \in \mathbb{R} \times \Theta$, which is the same K-dimensional space only with a different parametersation. In particular, $\hat{\boldsymbol{\theta}}^{\mathrm{c}}$ is non-unique if and only if $\hat{\boldsymbol{\theta}}^{\mathrm{u}}$ is non-unique.

Since one can obtain the solution to the constrained objective by solving the unconstrained one and then reparameterising (and vice versa), we are free to assume without loss of generality that $w^T \bar{\boldsymbol{Y}} = 0$, so $\hat{\mu} = 0$, when solving the univariate problem, and will remark where we do this.

2.8.4 Computation time experiments

A small experiment was performed to demonstrate the runtimes one can expect in practice for the univariate problem. Note that this clustering is applied iteratively in the block coordinate descent procedure we propose to use in multivariate settings. We considered 3 settings: one with no signal, one with 2 true clusters and one with 5 true clusters. Independent

and identically distributed Gaussian noise was added to each of the subaverages. As in Section 2.6.2 the number of categories was increased by random splitting of the levels. Each of these tests were repeated 25 times, on a computer with a 3.2GHz processor. The results are shown in Figure 2.8.

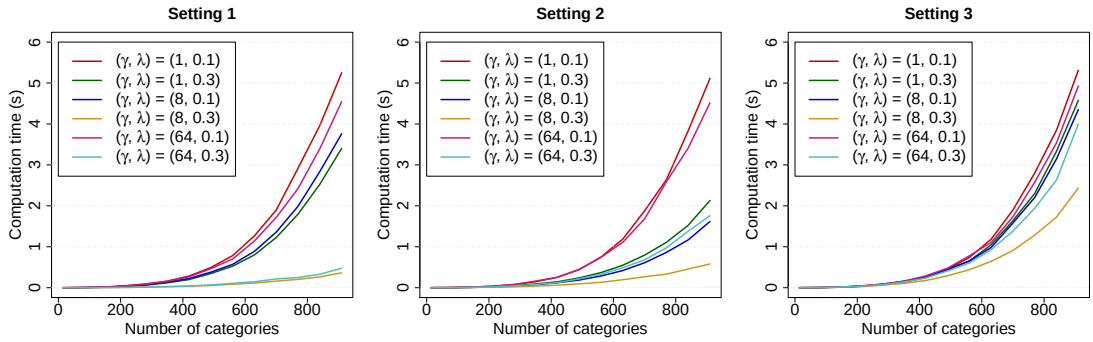


Figure 2.8 Computation times for solving the univariate problem.

2.8.5 Discretised algorithm

For very large-scale problems, speed can be improved if we only allow coefficients to take values in some fixed finite grid, rather than any real value. Below we describe how such an algorithm would approximately solve the univariate objective (2.12). We will use the unconstrained objective as discussed in Section 2.8.3. We would first fix L grid points $\vartheta_1 < \cdots < \vartheta_L$, and then proceed as described in Algorithm 2.3.

Algorithm 2.3 Discrete algorithm for computing approximate solution to (2.12)

for $l = 1, \ldots, L$ **do**
  Set $F_{\text{new}}(l) = \frac{1}{2} w_1 (\bar{Y}_1 - \vartheta_l)^2$
  Set $B(1, l) = l$
end for
for $k = 2, \ldots, K$ **do**
  Set $F_{\text{old}} = F_{\text{new}}$
  for $l = 1, \ldots, L$ **do**
    Set $B(k, l) = \arg\min_{l' \in \{1, \ldots, l\}} F_{\text{old}}(l') + \rho(\vartheta_l - \vartheta_{l'}) + \frac{1}{2} w_k (\bar{Y}_k - \vartheta_l)^2$
    Set $F_{\text{new}}(l) = F_{\text{old}}(B(k, l)) + \rho(\vartheta_l - \vartheta_{B(k,l)}) + \frac{1}{2} w_k (\bar{Y}_k - \vartheta_l)^2$
  end for
end for
Set $B^*(K) = \arg\min F_{\text{new}}$, and $\hat{\theta}_K = \vartheta_{B^*(K)}$
for $k = K - 1, \ldots, 1$ **do**
  Set $B^*(k) = B(k + 1, B^*(k + 1))$, and $\hat{\theta}_k = \vartheta_{B^*(k)}$
end for

This algorithm has the same basic structure to the approach we use in Section 2.3.1 for computing the exact global optimum. The difference is that now, instead of as in (2.14), we define f_k in the following way:

$$f_k(\theta_k) = \min_{\substack{(\theta_1,\ldots,\theta_{k-1})^T \in \{\vartheta_1,\ldots,\vartheta_L\}^{k-1} \\ \theta_1 \leq \cdots \leq \theta_{k-1} \leq \theta_k}} \left\{ \frac{1}{2} \sum_{l=1}^{k} w_l (\bar{Y}_l - \theta_l)^2 + \sum_{l=1}^{k-1} \rho(\theta_{l+1} - \theta_l) \right\}.$$

The objects F and B play analogous roles to f_k and b_k in Section 2.3.1. Since we restrict $\theta_k \in \{\vartheta_1, \ldots, \vartheta_L\}$, we only need to store the values that f_k takes at these L values; this is the purpose of the vector F in Algorithm 2.3. Similarly, the rows $B(k, \cdot)$ serve the same purpose as the functions b_k where, again, we only need to store L values corresponding to the different options for θ_k.

This algorithm returns the optimal solution $\hat{\boldsymbol{\theta}}$ to the objective where each of the coefficients are restricted to take values only in $\{\vartheta_1, \ldots, \vartheta_L\}$. We must ensure that the grid of values has fine enough resolution that interesting answers can be obtained, which requires L being sufficiently large. The number of clusters obtained by this approximate algorithm is bounded above by L, so this must not be chosen too small.

One can see that the computational complexity of this algorithm is linear in K, with a total of $O(KL^2)$ operations required. This is of course in addition to the $O(n)$ operations needed to compute $w_1, \ldots, w_K$ and $\bar{Y}_1, \ldots, \bar{Y}_K$ beforehand. In particular, choosing $L \lesssim \sqrt{K}$ guarantees that the complexity of this algorithm is at worst quadratic in K.

2.9 Proofs of results in Sections 2.2 & 2.3

Proof of Proposition 2.1. Assume, without loss of generality, that $\hat{\mu} = 0$. Suppose that there exists $l \neq k$ such that $\hat{\theta}_k = \hat{\theta}_l$. Without loss of generality we have that $\bar{Y}_k \neq \hat{\theta}_k$ (if $\bar{Y}_k = \hat{\theta}_k$ then $\bar{Y}_l \neq \hat{\theta}_l$ and it can be seen that $\hat{\theta}_{(1)} < \bar{Y}_l < \hat{\theta}_K$, in which case swap labels).

Now we construct $\tilde{\boldsymbol{\theta}}$ by setting $\tilde{\theta}_r = \hat{\theta}_r \wedge \bar{Y}_k$ for $r = 1, \ldots, k$, and $\tilde{\theta}_r = \hat{\theta}_r$ otherwise. We have $\ell(\hat{\mu}, \tilde{\boldsymbol{\theta}}) < \ell(\hat{\mu}, \hat{\boldsymbol{\theta}})$ and, by convexity of ρ, it follows that

$$\sum_{r=1}^{K-1} \rho(\tilde{\theta}_{(r+1)} - \tilde{\theta}_{(r)}) \leq \sum_{r=1}^{K-1} \rho(\hat{\theta}_{(r+1)} - \hat{\theta}_{(r)}).$$

This gives the conclusion $Q(\tilde{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\theta}$. □

Proof of Proposition 2.2. Suppose, for a contradiction, that $\hat{\theta}_k < \hat{\theta}_l$. Then at least one of the following must be true:

$$\left|\hat{\mu} + \hat{\theta}_k - \bar{Y}_k\right| > \left|\hat{\mu} + \hat{\theta}_l - \bar{Y}_k\right| \tag{2.32}$$

$$\left|\hat{\mu} + \hat{\theta}_l - \bar{Y}_l\right| > \left|\hat{\mu} + \hat{\theta}_k - \bar{Y}_l\right|. \tag{2.33}$$

Let $\tilde{\boldsymbol{\theta}}$ be defined as follows. Set $\tilde{\theta}_r = \hat{\theta}_r$ for all $r \neq k, l$. If (2.32) holds set $\tilde{\theta}_k = \hat{\theta}_l$ and if (2.33) holds set $\tilde{\theta}_l = \hat{\theta}_k$. Observe that

$$\sum_{r=1}^{n} \rho(\hat{\theta}_{(r+1)} - \hat{\theta}_{(r)}) \geq \sum_{r=1}^{n} \rho(\tilde{\theta}_{(r+1)} - \tilde{\theta}_{(r)})$$

and that the squared loss of $\tilde{\boldsymbol{\theta}}$ is strictly smaller than the squared loss of $\hat{\boldsymbol{\theta}}$, thus contradicting optimality of $\hat{\boldsymbol{\theta}}$. □

Proof of Proposition 2.3. In this proof we consider the unconstrained formulation of the objective (2.31) discussed in Section 2.8.3. Suppose that $(\bar{Y}_k)_{k=1}^K$ is such that there are two distinct solutions to (2.12), $\hat{\boldsymbol{\theta}}^{(1)} \neq \hat{\boldsymbol{\theta}}^{(2)}$. Let us assume that the levels are indexed such that $\bar{Y}_1 \leq \cdots \leq \bar{Y}_K$. Define $k^* = \max\{k : \hat{\theta}_k^{(1)} \neq \hat{\theta}_k^{(2)}\}$ to be the largest index at which the two solutions take different values and note that we must have $\hat{\theta}_1^{(r)} \leq \cdots \leq \hat{\theta}_K^{(r)}$.

First consider the case where $k^* < K$. Then

$$S_r = \{k : \hat{\theta}_k^{(r)} = \hat{\theta}_{k^*+1}^{(r)}\} \subseteq \{k^* + 1, k^* + 2, \ldots, K\},$$

for $r = 1, 2$. We now argue that we must have $\hat{\theta}_{k^*+1}^{(1)} = \hat{\theta}_{k^*+1}^{(2)} = t^* \geq (\hat{\theta}_{k^*}^{(1)} \vee \hat{\theta}_{k^*}^{(2)}) + \gamma\lambda$. Indeed, suppose not, and suppose that without loss of generality $\hat{\theta}_{k^*}^{(2)} > \hat{\theta}_{k^*}^{(1)}$. Fix $r \in \{1, 2\}$. The directional derivative of the objective in the direction of the binary vector with ones at the indices given by S_r and zeroes elsewhere evaluated at $\hat{\boldsymbol{\theta}}^{(r)}$ must be 0. But comparing these for $r = 1, 2$, we see they are identical except for the term $\rho'(\theta_{k^*+1} - \hat{\theta}_{k^*}^{(r)})$, which will be strictly larger for $r = 2$, giving a contradiction. This then implies that both $\hat{\theta}_{k^*}^{(1)}$ and $\hat{\theta}_{k^*}^{(2)}$ must minimise f_{k^*} over $\theta \leq t^* - \gamma\lambda$ since the full objective value is

$$Q(\hat{\boldsymbol{\theta}}^{(r)}) = f_{k^*}(\hat{\theta}_{k^*}^{(r)}) + \frac{1}{2}\gamma\lambda^2 + (\text{terms featuring only index } k^* + 1 \text{ or higher})$$

for $r = 1, 2$. We also have that when $k^* = K$, both $\hat{\theta}_{k^*}^{(1)}$ and $\hat{\theta}_{k^*}^{(2)}$ must minimise f_{k^*}.

Using the functions g_{k-1} as defined in (2.15), we have the simple relationship that $g_{k-1}(\theta_k) = f_k(\theta_k) - \frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$. In particular, properties (i) and (iii) of Lemma 2.4 hold with f_k replaced by g_{k-1}. These can be characterised as $g_{k-1}(\theta_k) = \breve{q}_{k,r}(\theta_k)$ for $\theta_k \in I_{k,r}$, where $I_{k,r}$ are the intervals associated with f_k and $\breve{q}_{k,r}(\theta_k) = q_{k,r}(\theta_k) - \frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$. Note

that for each r, $\check{q}_{k,r}$ depends on the values of $\bar{Y}_1, \ldots, \bar{Y}_{k-1}$ but not that of $\bar{Y}_k$ (observe that $q_{k,r}(\theta_k)$ includes a term $\frac{1}{2}w_k(\bar{Y}_k - \theta_k)^2$; see (2.13)).

Now as $\hat{\theta}^{(1)}_{k^*} \leq \hat{\theta}^{(1)}_{k^*+1} - \gamma\lambda$ and $\hat{\theta}^{(2)}_{k^*} \leq \hat{\theta}^{(2)}_{k^*+1} - \gamma\lambda$ (if $k^* < K$), by Lemma 2.4 (iii) both must be local minima of f_{k^*}, and we have that there must exist distinct $r_1 \neq r_2$ such that $\hat{\theta}^{(1)}_{k^*} \in I_{k^*,r_1}$ and $\hat{\theta}^{(2)}_{k^*} \in I_{k^*,r_2}$. Let

$$\check{q}_{k^*,r_1}(x) = a_1x^2 + b_1x + c_1,$$
$$\check{q}_{k^*,r_2}(x) = a_2x^2 + b_2x + c_2.$$

Since $\hat{\theta}^{(1)}_{k^*}$ must be the minimum of $\check{q}_{k^*,r_1}(\theta_{k^*}) + \frac{1}{2}w_{k^*}(\bar{Y}_{k^*} - \theta_{k^*})^2$ (and similarly for $\hat{\theta}^{(2)}_{k^*}$), we must have that

$$\min_x \left\{ a_1x^2 + b_1x + c_1 + \frac{1}{2}w_{k^*}(\bar{Y}_{k^*} - x)^2 \right\} = \min_x \left\{ a_2x^2 + b_2x + c_2 + \frac{1}{2}w_{k^*}(\bar{Y}_{k^*} - x)^2 \right\}$$
$$\implies c_1 - \frac{(b_1 - w_{k^*}\bar{Y}_{k^*})^2}{4a_1 + 2w_{k^*}} = c_2 - \frac{(b_2 - w_{k^*}\bar{Y}_{k^*})^2}{4a_2 + 2w_{k^*}}. \tag{2.34}$$

This is a quadratic equation in $\bar{Y}_{k^*}$, so there are at most two values for which (2.34) holds. Considering all pairs r_1, r_2, we see that in order for there to exist two solutions $\hat{\boldsymbol{\theta}}^{(1)} \neq \hat{\boldsymbol{\theta}}^{(2)}$, $\bar{Y}_{k^*}$ must take values in a set of size at most $c(K)$, for some function $c : \mathbb{N} \to \mathbb{N}$.

Now let

$$\mathcal{S} = \{(\bar{Y}_k)_{k=1}^K : \text{ the minimiser of the objective is not unique}\} \subseteq \mathbb{R}^K.$$

What we have shown, is that associated with each element $(\bar{Y}_k)_{k=1}^K \in \mathcal{S}$, there is at least one k^* such that

$$|\{(\bar{Y}'_k)_{k=1}^K \in \mathcal{S} : \bar{Y}'_k = \bar{Y}_k \text{ for all } k \neq k^*\}|$$

is bounded above by $c(K)$. Now for each $j = 1, \ldots, K$, let $\mathcal{S}_j$ be the set of $(\bar{Y}_k)_{k=1}^K \in \mathcal{S}$ for which the there exists a k^* with the property above and $k^* = j$. Note that $\cup_j \mathcal{S}_j = \mathcal{S}$. Now $\mathcal{S}_j \subset \mathbb{R}^K$ has Lebesgue measure zero as a finite union of graphs of measurable functions $f : \mathbb{R}^{K-1} \to \mathbb{R}$. Thus $\mathcal{S}$ has Lebesgue measure zero. □

Proof of Lemma 2.4. Assume, without loss of generality, that $\hat{\mu} = 0$. We proceed inductively, assuming that the properties (i) and (iii) hold for f_k, and (ii) holds for b_{k+1}. Additionally we include in our inductive hypothesis that for all x, $f'_k(x_-) \geq f'_k(x_+)$, where we define $f'_k(x_-)$ and $f'_k(x_+)$ to be the left-derivative and right-derivative of f_k at x, respectively. We note that these trivially hold for the base case f_1, and the case b_2 can be checked by direct calculation.

We first prove (i), that f_{k+1} is continuous, coercive, and piecewise quadratic and with finitely many pieces. We then show that $f'_{k+1}(x_-) \geq f'_{k+1}(x_+)$ for all x, which allows us to show that (iii) holds for f_{k+1}. Finally, we use these results to show that (ii) holds for b_{k+2}.

We now show that f_{k+1} is coercive and continuous. Clearly $g_k(x) \geq \min_{y \leq x} f_k(y)$, so it follows that $g_k(x) \to \infty$ as $x \to -\infty$ as f_k is coercive. Furthermore g_k is bounded from below as f_k is coercive and continuous. Thus since $f_{k+1}(x) = g_k(x) + \frac{1}{2} w_{k+1}(\bar{Y}_{k+1} - x)^2$, it follows that f_{k+1} is coercive. Next as $g_k(x) = \min_{y \leq x}\{f_k(y) + \rho(y - x)\}$, and f_k and ρ are continuous, it follows that g_k is continuous and therefore that f_{k+1} is continuous.

To see why f_{k+1} is piecewise quadratic with finitely many pieces, we observe that it can be written $f_{k+1}(x) = f_k(b_{k+1}(x)) + \rho(x - b_{k+1}(x)) + \frac{1}{2} w_{k+1}(\bar{Y}_{k+1} - x)^2$. We have by our inductive hypothesis that f_k is piecewise quadratic and $b_{k+1}(x)$ is piecewise linear, both with finitely many pieces. Since the composition of a piecewise linear function inside a piecewise quadratic function is piecewise quadratic, the remainder of (i) is shown.

We now turn our attention to (iii), and define for $x \in \mathbb{R}$:

$$y_*(x) = \operatorname*{sarg\,min}_{y \leq x} \{f_k(y) + \rho(x - y)\},$$
$$y^*(x) = \operatorname*{sarg\,min}_{y \leq x} \{f_{k+1}(y) + \rho(x - y)\}.$$

We will first show that $f'_{k+1}(x_+) \leq f'_{k+1}(x_-)$ for all $x \in \mathbb{R}$. Suppose that we are increasing x and we have reached a point where $g_k(x)$ is not differentiable (that is, the left-derivative and the right-derivative do not match). By assumption (ii) for b_{k+1}, we can assume that there is some window $\delta > 0$ such that $y_*(t)$ is linear for $t \in (x - \delta, x)$, say $y_*(t) = \alpha + \beta t$.

In order to proceed with the following argument, we must show that for sufficiently small $\epsilon > 0$, we have $\alpha + \beta(x + \epsilon) \leq x + \epsilon$. If $\alpha + \beta x < x$, this is immediate. Therefore it remains to consider the case $\alpha + \beta x = x$, for which we show that we must have $\alpha = 0$ and $\beta = 1$, i.e $y_*(t) = t$ for $t \in (x - \delta, x)$. This follows from the observation that if $y_*(t) < t$, then for all $t_1 > t$ we have $y_*(t_1) \notin (y_*(t), t]$. Indeed, suppose not, then

$$\begin{aligned} & f_k(y_*(t_1)) + \rho(t_1 - y_*(t_1)) < f_k(y_*(t)) + \rho(t_1 - y_*(t)) \\ \implies \quad & f_k(y_*(t_1)) + \rho(t - y_*(t_1)) < f_k(y_*(t)) + \rho(t_1 - y_*(t)) + \rho(t - y_*(t_1)) - \rho(t_1 - y_*(t_1)) \\ & \leq f_k(y_*(t)) + \rho(t - y_*(t)), \end{aligned}$$

contradicting the definition of $y_*(t)$. The last line uses $\rho(t_1 - y_*(t)) - \rho(t_1 - y_*(t_1)) \leq \rho(t - y_*(t)) - \rho(t - y_*(t_1))$, which follows from concavity of ρ and $y_*(t) < y_*(t_1) \leq t < t_1$.

With this established, we have that:

$$\begin{aligned} g_k(x - \epsilon) &= f_k(\alpha + \beta(x - \epsilon)) + \rho(x - \epsilon - (\alpha + \beta(x - \epsilon))) \\ g_k(x + \epsilon) &= f_k(y_*(x + \epsilon)) + \rho(x + \epsilon - y_*(x + \epsilon)) \\ &\leq f_k(\alpha + \beta(x + \epsilon)) + \rho(x + \epsilon - (\alpha + \beta(x + \epsilon))). \end{aligned}$$

Note that f_k has both left-derivatives and right-derivatives at every point in $\mathbb{R}$. Suppose first that $\beta \geq 0$, and we observe that

$$g_k'(x_-) = \beta f_k'(y_*(x)_-) + (1-\beta)\rho'(x - y_*(x))$$

Then by the basic definition of the right-derivative,

$$\begin{aligned}
g_k'(x_+) &= \lim_{\epsilon \to 0^+} \frac{f_k(y_*(x+\epsilon)) + \rho(x+\epsilon - y_*(x+\epsilon)) - f_k(y_*(x)) - \rho(x - y_*(x))}{\epsilon} \\
&\leq \lim_{\epsilon \to 0^+} \frac{1}{\epsilon}\Big[f_k(\alpha + \beta(x+\epsilon)) + \rho(x+\epsilon - (\alpha + \beta(x+\epsilon))) \\
&\qquad\qquad - f_k(\alpha + \beta x) - \rho(x - (\alpha + \beta x))\Big] \\
&= \beta f_k'(y_*(x)_+) + (1-\beta)\rho'(x - y_*(x)) \\
&= g_k'(x_-) + \beta(f_k'(y_*(x)_+) - f_k'(y_*(x)_-)) \\
&\leq g_k'(x_-),
\end{aligned}$$

where the last inequality follows from our inductive hypothesis that $f_k'(y_+) \leq f_k'(y_-)$ for all $y \in \mathbb{R}$. An analogous argument shows that the same conclusion holds when $\beta < 0$.

Now we use this to prove the claim. Because there are no points of f_{k+1} at which the left-derivative is less than the right-derivative, without loss of generality we claim that f_{k+1} is differentiable at $y^*(x)$ for all x, unless $y^*(x) = x$. Indeed, suppose not, then we have that $f_{k+1}'(y^*(x)_-) > f_{k+1}'(y^*(x)_+)$ and necessarily that defining $h(y) = f_{k+1}(y) + \rho(x-y)$, we have $0 \in \partial h(y^*(x))$. But since $h(y^*(x)_+) < h(y^*(x)_-)$, we contradict the optimality of $y^*(x)$ as this point is in fact a local maximum.

We finally consider claim (ii). By (iii), we have that for every point x, $y^*(x)$ is either x or at the minimum of one of the quadratic pieces of $f_{k+1}(\cdot) + \rho(x - \cdot)$. In either case, we have that $y^*(x)$ is linear in x and thus $f_{k+1}(y^*(x)) + \rho(x - y^*(x))$ is quadratic in x. We can define $g_{k+1}(x)$ pointwise as the minimum of this finite set of quadratic functions of x, whose expressions are given in Section 2.8.1. Importantly, the coefficients in the linear expression $y^*(x)$ of x depend only on which of these functions is the minimum at x. As the number of intersections between elements in this set of quadratic functions is bounded above by twice the square of the size of the set, we can conclude that $b_{k+2}(x)$ is piecewise linear and with a finite number of pieces, thus concluding the proof. □

2.10 Proofs of results in Section 2.4

2.10.1 Proof of Theorem 2.5

The proof of Theorem 2.5 requires a number of auxiliary lemmas, which can be found in Section 2.10.1.

Let us define $R_i = Y_i - \hat{\mu}$ for $i = 1, \ldots, n$, and $\bar{R}_k = \frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i=k\}} R_i$ for $k = 1, \ldots, K$. Note that

$$R_i = \sum_{k=1}^K \mathbb{1}_{\{X_i=k\}} \theta_k^0 + (P\varepsilon)_i$$

where $P = I - \mathbf{1}\mathbf{1}^T/n$.

For each $k = 1, \ldots, K$, we define the event

$$\Lambda_k = \left\{ \left| \frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i=k\}} (P\varepsilon)_i \right| < \frac{1}{2}\sqrt{\eta\gamma_* s}\lambda \right\}.$$

By a union bound, we have that $\mathbb{P}(\cap_{k=1}^K \Lambda_k) \geq 1 - \sum_{k=1}^K \mathbb{P}(\Lambda_k^c)$. Now observe we can write

$$\frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i=k\}} (P\varepsilon)_i = {v^{(k)}}^T P\varepsilon,$$

where we define $v^{(k)} \in \mathbb{R}^n$ by $v_i^{(k)} = \frac{1}{n_k} \mathbb{1}_{\{X_i=k\}}$. Since P is an orthogonal projection matrix, we have that $\|Pv^{(k)}\|_2 \leq \|v^{(k)}\|_2 = \frac{1}{\sqrt{n_k}}$. It follows that ${v^{(k)}}^T P\varepsilon$ is sub-Gaussian with parameter $\sigma/\sqrt{n_k}$. Applying the standard sub-Gaussian tail bound, we obtain

$$\begin{aligned}
\mathbb{P}(\Lambda_k^c) &= \mathbb{P}\left(\left| \frac{1}{n_k} \sum_{i=1}^n \mathbb{1}_{\{X_i=k\}} (P\varepsilon)_i \right| \geq \frac{1}{2}\sqrt{\eta\gamma_* s}\lambda \right) \\
&\leq 2\exp\left(-\frac{n w_k \eta\gamma_* s\lambda^2}{8\sigma^2} \right),
\end{aligned}$$

where recall that $w_k = n_k/n$. Therefore, we have that

$$\mathbb{P}\left(\cap_{k=1}^K \Lambda_k\right) \geq 1 - 2\sum_{k=1}^K \exp\left(-\frac{n w_k \eta\gamma_* s\lambda^2}{8\sigma^2} \right) \geq 1 - 2\exp\left(-\frac{n w_{\min} \eta\gamma_* s\lambda^2}{8\sigma^2} + \log(K) \right). \tag{2.35}$$

In the following we work on the intersection $\Lambda = \cap_{k=1}^K \Lambda_k$. This entails that for each k, $|\bar{R}_k - \theta_k^0| < \sqrt{\eta\gamma_* s}\lambda/2$. We now relabel indices such that $\bar{R}_1 \leq \cdots \leq \bar{R}_K$, and so from Proposition 2.2 that $\hat{\theta}_1 \leq \cdots \leq \hat{\theta}_K$. Since our assumption (2.24) implies $\Delta(\boldsymbol{\theta}^0) \geq \sqrt{\eta\gamma_* s}\lambda$, it

follows that on Λ the observed ordering is consistent with the ordering of the true coefficients, i.e. there exist $0 = k_0 < k_1 < \cdots < k_s = K$ such that

$$\theta^0_1 = \cdots = \theta^0_{k_1} < \theta^0_{k_1+1} = \cdots = \theta^0_{k_2} < \cdots < \theta^0_{k_{s-1}+1} = \cdots = \theta^0_{k_s}. \tag{2.36}$$

Indeed, observe that for $j = 1, \ldots, s-1$, we have by the triangle inequality and (2.24), the stronger property that

$$\begin{aligned} \bar{R}_{k_j+1} - \bar{R}_{k_j} &> 3\left(1 + \frac{\sqrt{2}}{\eta}\right)\sqrt{\gamma\gamma^*}\lambda - \sqrt{\eta\gamma_* s}\lambda \\ &> \gamma\lambda + 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + 2\sqrt{\eta\gamma_* s}\lambda. \end{aligned} \tag{2.37}$$

Our optimisation objective is therefore

$$\hat{\boldsymbol{\theta}} \in \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{R}_k - \theta_k)^2 + \sum_{k=1}^{K} \rho(\theta_{k+1} - \theta_k) \right\}. \tag{2.38}$$

Since $\bar{R}_{k_j} - \bar{R}_{k_{j-1}+1} < \sqrt{\eta\gamma_* s}\lambda$ for $j = 1, \ldots, s$, it follows from Lemma 2.8 that $\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j} \geq \gamma\lambda$ for $j = 1, \ldots, s-1$, so

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) &= \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{R}_k - \hat{\theta}_k)^2 + \sum_{k=1}^{K-1} \rho(\hat{\theta}_{k+1} - \hat{\theta}_k) \\ &= \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{R}_k - \hat{\theta}_k)^2 + \sum_{j=1}^{s}\sum_{k=k_{j-1}+1}^{k_j-1} \rho(\hat{\theta}_{k+1} - \hat{\theta}_k) + \frac{s-1}{2}\gamma\lambda^2 \qquad (2.39) \\ &= \min_{\boldsymbol{\theta}\in\mathbb{R}^K} \left\{ \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{R}_k - \theta_k)^2 + \sum_{j=1}^{s}\sum_{k=k_{j-1}+1}^{k_j-1} \rho(\theta_{k+1} - \theta_k) \right\} + \frac{s-1}{2}\gamma\lambda^2. \qquad (2.40) \end{aligned}$$

Observe that we can have $k_{j-1} + 1 > k_j - 1$ for some j, in which case we take the sum over that range to be zero. Note that (2.40) can be optimised over $(\theta_{k_{j-1}+1}, \ldots, \theta_{k_j})$ separately for each $j = 1, \ldots, s$. If $s = 1$, i.e. the true signal is zero, then the result follows from Lemma 2.10. Now we see what happens when $s > 1$.

Without loss of generality, consider $j = 1$ and note that if $k_1 = 1$ it is immediate that $\hat{\theta}_1 = \hat{\theta}^0_1$. Hence, we can assume that $k_1 > 1$. We note that $\hat{\theta}^0_1 = \sum_{k=1}^{k_1} w_k\bar{R}_k/w^0_1$, where we define $w^0_k = n^0_k/n$. We see that our goal is to compute

$$\begin{aligned} &\operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{k_1}} \left\{ \frac{1}{2}\sum_{k=1}^{k_1} w_k(\bar{R}_k - \theta_k)^2 + \sum_{k=1}^{k_1-1} \rho(\theta_{k+1} - \theta_k) \right\} \\ &= \hat{\theta}^0_1\mathbf{1} + \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^{k_1}} \left\{ \frac{1}{2}\sum_{k=1}^{k_1} w_k(\tilde{R}_k - \theta_k)^2 - \sum_{k=1}^{k_1-1} \rho(\theta_{k+1} - \theta_k) \right\}, \end{aligned} \tag{2.41}$$

where $\mathbf{1} \in \mathbb{R}^{k_1}$ is a vector of ones and $\tilde{R}_k = \bar{R}_k - \hat{\theta}_1^0$ for $k = 1, \ldots, k_1$. Note that we subtract $\hat{\theta}_1^0$ to ensure that

$$\sum_{k=1}^{k_1} w_k \tilde{R}_k = 0,$$

as required for application of Lemma 2.10. We have by assumption that for $k \in 1, \ldots, k_1$, $|\tilde{R}_k| \leq \sqrt{\eta\gamma_* s}\lambda/2 \leq (2 \wedge \sqrt{w_1^0\gamma})\lambda/w_1^0$. Thus, Lemma 2.10 can be applied with $\check{w} = w_1^0$ and it follows that $\hat{\theta}_k = \hat{\theta}_1^0$ for $k = 1, \ldots, k_1$. □

Auxiliary lemmas

Here we prove a number of results required to obtain conditions for recovering the oracle least squares estimate in the univariate case. Lemma 2.10 gives conditions for recovery of the true solution, in the case where there is zero signal. Lemmas 2.8 and 2.9 ensure that the true levels are far enough apart that they can be separated. Once we have this separation, we apply Lemma 2.10 on each of the levels to obtain the solution.

Lemma 2.7. *Consider the optimisation problem*

$$x^* = \operatorname*{arg\,min}_{x \geq 0} \left\{ \frac{\kappa}{2}(2\tau - x)^2 + \rho(x) \right\},$$

where $\tau > 0$ and $\kappa \in (0, 1]$. Suppose further that $\tau < (1 \wedge \sqrt{\kappa\gamma})\lambda/2\kappa$. Then $x^* = 0$ is the unique optimum.

Proof. We first observe that

$$x^* = \operatorname*{arg\,min}_{x \geq 0} \left\{ \frac{\kappa}{2}(2\tau - x)^2 + \rho_{\gamma,\lambda}(x) \right\} = \operatorname*{arg\,min}_{x \geq 0} \left\{ \frac{1}{2}(2\tau - x)^2 + \rho_{\kappa\gamma,\lambda/\kappa}(x) \right\}.$$

For convenience, we define $F(x) = (2\tau - x)^2/2 + \rho_{\kappa\gamma,\lambda/\kappa}(x)$. It now suffices to show that F is uniquely minimised at 0 provided $\tau < (1 \wedge \sqrt{\kappa\gamma})\lambda/2\kappa$. We can clearly see that $x^* \in [0, 2\tau]$. Equation (2.3) of Breheny and Huang (2011) gives the result when $\kappa\gamma \geq 1$.

When $\kappa\gamma < 1$, we see that any stationary point of F in $[0, \gamma\lambda \wedge 2\tau]$ must be a maximum, since on this interval $F(x)$ is a quadratic function with a negative coefficient of x^2. Therefore its minimum over $[0, \gamma\lambda]$ is attained at either $x = 0$ or $x = \gamma\lambda \wedge 2\tau$. If $2\tau \leq \gamma\lambda$, then it suffices to check that $F(0) < F(2\tau)$. This holds if and only if $\tau < \gamma\lambda/(\gamma\kappa + 1)$, but since we are assuming $\tau \leq \gamma\lambda/2$ and $\kappa\gamma < 1$, this is always satisfied.

If $\gamma\lambda < 2\tau$, then we can see that the minimum of F over $[\gamma\lambda, 2\tau]$ will be attained at exactly 2τ. Thus, here it also suffices to check $F(0) < F(2\tau)$, which holds if and only if $\tau < \sqrt{\gamma/\kappa}\lambda/2$. The final bound $\tau < (1 \wedge \sqrt{\kappa\gamma})\lambda/2\kappa$ follows from combining the results for these cases. □

The following is a deterministic result to establish separation between groups of coefficients.

Lemma 2.8. *Consider the setup of Theorem 2.5, and assume that $\hat{\mu} = 0$. Suppose that $\bar{Y}_1 \leq \cdots \leq \bar{Y}_K$, and that for $j = 1, \ldots, s$ we have*

$$\bar{Y}_{k_j} - \bar{Y}_{k_{j-1}+1} < \sqrt{\eta\gamma_* s}\lambda, \tag{2.42}$$

where k_j and k_{j-1} are as defined in (2.36)*. Suppose further that for $j = 1, \ldots, s-1$,*

$$\bar{Y}_{k_j+1} - \bar{Y}_{k_j} \geq \gamma\lambda + 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + 2\sqrt{\eta\gamma_* s}\lambda. \tag{2.43}$$

Then for $j = 1, \ldots, s$, we have $\bar{Y}_{k_{j-1}+1} \leq \hat{\theta}_{k_{j-1}+1} \leq \hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$.

Proof. For convenience, within this lemma we define $\zeta = \sqrt{\eta\gamma_* s}\lambda$. Recall that the objective function which $\hat{\boldsymbol{\theta}}$ optimises takes the form

$$Q(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \theta_k)^2 + \sum_{k=1}^{K-1} \rho(\theta_{k+1} - \theta_k).$$

We first claim that $\hat{\theta}_k \in [\bar{Y}_1, \bar{Y}_K]$ for $k = 1, \ldots, K$. To see this, suppose that this is not the case and define $\check{\boldsymbol{\theta}}$ by projecting $\hat{\boldsymbol{\theta}}$ onto $[\bar{Y}_1, \bar{Y}_K]^K$ (i.e. $\check{\theta}_k = \bar{Y}_K \wedge (\bar{Y}_1 \vee \hat{\theta}_k)$ for $k = 1, \ldots, K$). The penalty contribution from $\check{\boldsymbol{\theta}}$ is no larger than that of $\hat{\boldsymbol{\theta}}$, and the loss contribution is strictly smaller, so we obtain the contradiction $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$.

We now proceed to show that for $j = 1, \ldots, s-1$, we have $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$ and $\hat{\theta}_{k_j+1} \geq \bar{Y}_{k_j+1}$. We prove the first of these sets of inequalities, since the second follows similarly by considering the problem with $-\hat{\boldsymbol{\theta}}$, $-\bar{Y}$ and reversing the indices. Suppose, for contradiction, that there exists some j in $\{1, \ldots, s-1\}$ with $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$. Let this j be minimal, such that for all $l < j$ we have $\hat{\theta}_{k_l} \leq \bar{Y}_{k_l}$.

Next define l_1 to be the maximal element of $\{k_{j-1}+1, \ldots, k_j - 1\}$ such that $\hat{\theta}_{l_1} \leq \bar{Y}_{k_j}$. Similarly, we define $l_2 \in \{k_j+1, \ldots, k_{j+1}\}$ to be minimal such that $\hat{\theta}_{l_2} \geq \bar{Y}_{k_j+1}$. The existence of l_1 and l_2 is guaranteed by Lemma 2.9.

We note that for $l = l_1+1, \ldots, k_j$, $\hat{\theta}_l = \hat{\theta}_{k_j}$ and hence $(\bar{Y}_l - \hat{\theta}_l)^2 \geq (\bar{Y}_{k_j} - \hat{\theta}_l)^2 = (\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2$. This can be shown by contradiction, as in (2.55). For such l, we have from optimality of $\hat{\theta}$ that $\bar{Y}_l - \hat{\theta}_{l_1} \geq \hat{\theta}_{k_j} - \bar{Y}_l$ (otherwise one could improve the objective by setting $\hat{\theta}_{l_1} = \hat{\theta}_l$) which implies that $\hat{\theta}_{l_1} < \bar{Y}_l$. From this it follows that $(\bar{Y}_l - \hat{\theta}_{l_1})^2 \leq (\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2$, since $\hat{\theta}_{l_1} < \bar{Y}_l \leq \bar{Y}_{k_j}$.

Similarly, if $l_2 > k_j + 1$, then for $l = k_j + 1, \ldots, l_2 - 1$ we have $\hat{\theta}_l = \hat{\theta}_{k_j+1}$ and hence $(\bar{Y}_l - \hat{\theta}_l)^2 \geq (\bar{Y}_{k_j+1} - \hat{\theta}_l)^2 = (\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1})^2$. For such l, it follows that $\hat{\theta}_{l_2} > \bar{Y}_l$ and therefore that $(\bar{Y}_l - \hat{\theta}_{l_2})^2 \leq (\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2$.

Now, we define

$$\tilde{w}_{k_j} = \sum_{l \le k_j \,:\, \hat{\theta}_l = \hat{\theta}_{k_j}} w_l$$

$$\text{and, if } l_2 > k_j + 1, \qquad \tilde{w}_{k_j+1} = \sum_{l \ge k_j+1 \,:\, \hat{\theta}_l = \hat{\theta}_{k_j+1}} w_l.$$

We also define $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^K$ according to

$$\tilde{\theta}_l = \begin{cases} \hat{\theta}_l \wedge \hat{\theta}_{l_1} & \text{for } l \le k_j \\ \hat{\theta}_l \vee \hat{\theta}_{l_2} & \text{for } l > k_j. \end{cases}$$

We note that by assumption, both $\tilde{w}_{k_j} < 1/\eta s$ and $\tilde{w}_{k_j+1} < 1/\eta s$. We now consider two cases: (A) where $l_2 = k_j + 1$, so $\hat{\theta}_{k_j+1} \ge \bar{Y}_{k_j+1}$, and (B) where $l_2 > k_j + 1$, so $\hat{\theta}_{k_j+1} < \bar{Y}_{k_j+1}$.

We first consider case (A), where the penalty terms between l_1 and l_2 in $Q(\hat{\boldsymbol{\theta}})$ are

$$\sum_{l=l_1}^{l_2-1} \rho(\hat{\theta}_{l+1} - \hat{\theta}_l) = \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}).$$

Thus,

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) = & \sum_{l \le k_j \,:\, \hat{\theta}_l = \hat{\theta}_{k_j}} \frac{w_l}{2}(\bar{Y}_l - \hat{\theta}_l)^2 - \sum_{l \le k_j \,:\, \hat{\theta}_l = \hat{\theta}_{k_j}} \frac{w_l}{2}(\bar{Y}_l - \hat{\theta}_{l_1})^2 \\ & + \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \\ \ge & \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 \\ & + \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \qquad (2.44) \\ \ge & \inf_{\bar{Y}_{k_j} < a \le \hat{\theta}_{l_2}} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 \right. \\ & \left. + \rho(\hat{\theta}_{l_2} - a) + \rho(a - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \right\}. \qquad (2.45) \end{aligned}$$

We specify the infimum in (2.47) because $(\bar{Y}_{k_j}, \hat{\theta}_{l_2}]$ is not closed, and let (a_m) be a convergent sequence in $(\bar{Y}_{k_j}, \hat{\theta}_{l_2}]$ whose limit attains this infimum. We define $a^* = \lim_{m\to\infty} a_m$.

By assumption (2.43), at least one of $(a^* - \hat{\theta}_{l_1})$ and $(\hat{\theta}_{l_2} - a^*)$ is greater than or equal to $\gamma\lambda$. Here, we use that the separation (2.43) $\ge 2\gamma\lambda$. If $\hat{\theta}_{l_2} - a^* \ge \gamma\lambda$ then we denote this case

(A1) and (2.45) becomes

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \inf_{\bar{Y}_{k_j} < a \leq \hat{\theta}_{l_2} - \gamma\lambda} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(a - \hat{\theta}_{l_1}) \right\} \tag{2.46}$$

$$\geq \min_{\hat{\theta}_{l_1} \leq \tilde{a} \leq \hat{\theta}_{l_2} - \gamma\lambda} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}) \right\}. \tag{2.47}$$

We define $\tilde{a}^*$ to be the minimiser over $\tilde{a}$ of (2.47). We can observe that since $\bar{Y}_{k_j} - \hat{\theta}_{l_1} < \zeta$ and $\zeta < (1 \wedge \sqrt{\gamma\tilde{w}_{k_j}})\lambda/\tilde{w}_{k_j}$, we have $\bar{Y}_{k_j} - \hat{\theta}_{l_1} < (1 \wedge \sqrt{\gamma\tilde{w}_{k_j}})\lambda/\tilde{w}_{k_j}$. Thus, we have by Lemma 2.7 that the uniquely optimal $\tilde{a}^* = \hat{\theta}_{l_1}$. This gives that the value of (2.47) is zero.

It is straightforward to see from (2.46) that $a^* = \bar{Y}_{k_j}$ must be the unique limit of (a_m). As we have assumed that $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$ and the infimum is not attained in $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$, the inequality in line (2.46) can be made strict. It follows that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

Thus, it remains for us to consider the case where $\hat{\theta}_{l_2} - a^* < \gamma\lambda$, which implies that $a^* - \hat{\theta}_{l_1} \geq \gamma\lambda$. We denote this case (A2). Now, from (2.45) we can obtain

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{\hat{\theta}_{l_2} - \gamma\lambda < \tilde{a} \leq \hat{\theta}_{l_2}} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\hat{\theta}_{l_2} - \tilde{a}) \right\}. \tag{2.48}$$

The objective is piecewise quadratic (and continuously differentiable), with two pieces: $[\hat{\theta}_{l_1}, \hat{\theta}_{l_2} - \gamma\lambda]$ and $(\hat{\theta}_{l_2} - \gamma\lambda, \hat{\theta}_{l_2}]$. On the first region, the objective is a convex quadratic with minimum at $\bar{Y}_{k_j} \in [\hat{\theta}_{l_1}, \hat{\theta}_{l_2} - \gamma\lambda]$.

By the assumption that $a^* > \hat{\theta}_{l_2} - \gamma\lambda$, we know that the objective must be concave on $(\hat{\theta}_{l_2} - \gamma\lambda, \hat{\theta}_{l_2}]$. It is clear that the derivative of the objective at $\hat{\theta}_{l_2} - \gamma\lambda$ is positive. Hence, if $\tilde{a}^* = \hat{\theta}_{l_2} - \gamma\lambda$, then the objective will take a strictly lower value at some $\tilde{a}^* \in (\hat{\theta}_{l_2} - \gamma\lambda - \epsilon, \hat{\theta}_{l_2} - \gamma\lambda)$ (for some small $\epsilon > 0$), contradicting optimality of $\tilde{a}^*$. It therefore follows that $\tilde{a}^* = \hat{\theta}_{l_2}$.

With this knowledge, we can further simplify (2.48) to obtain

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_2})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 > 0.$$

The second inequality follows from $\bar{Y}_{k_j} - \hat{\theta}_{l_1} \leq \zeta$ and $\hat{\theta}_{l_2} - \bar{Y}_{k_j} > \zeta$. Hence, we obtain that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

We now we direct our attention towards case (B), where similarly to before we observe that the penalty contributions between l_1 and l_2 in $Q(\hat{\boldsymbol{\theta}})$ are

$$\sum_{l=l_1}^{l_2-1} \rho(\hat{\theta}_{l+1} - \hat{\theta}_l) = \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}).$$

Similarly to (2.44) in case (A), we obtain

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq & \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{k_j})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1})^2 \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 \\ & + \rho(\hat{\theta}_{l_2} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \end{aligned} \tag{2.49}$$

$$\begin{aligned} \geq & \inf_{\bar{Y}_{k_j} < a \leq b < \bar{Y}_{k_j+1}} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 \right. \\ & - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 \\ & \left. + \rho(\hat{\theta}_{l_2} - b) + \rho(b - a) + \rho(a - \hat{\theta}_{l_1}) - \frac{1}{2}\gamma\lambda^2 \right\}. \end{aligned} \tag{2.50}$$

We specify the infimum in (2.50) because $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$ is not closed and therefore a minimum may not exist. Let (a_m, b_m) be a convergent sequence in $(\bar{Y}_{k_j}, \bar{Y}_{k_j+1})$ whose limit achieves this infimum. We now define $(a^*, b^*) = \lim_{m\to\infty}(a_m, b_m)$. By assumption (2.43), we know that $\bar{Y}_{k_j+1} - \bar{Y}_{k_j} \geq 3\gamma\lambda$, which implies that $\hat{\theta}_{l_2} - \hat{\theta}_{l_1} \geq 3\gamma\lambda$. Thus, one of $\{(\hat{\theta}_{l_2} - b^*), (b^* - a^*), (a^* - \hat{\theta}_{l_1})\}$ must be at least $\gamma\lambda$.

We first consider if $b^* - a^* \geq \gamma\lambda$, and denote this case (B1). Here, (2.50) becomes

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq & \inf_{\bar{Y}_{k_j} < a \leq b < \bar{Y}_{k_j+1}} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 \right. \\ & \left. - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\hat{\theta}_{l_2} - b) + \rho(a - \hat{\theta}_{l_1}) \right\} \end{aligned} \tag{2.51}$$

$$\begin{aligned} = & \inf_{a \in (\bar{Y}_{k_j}, \bar{Y}_{k_j+1})} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - a)^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(a - \hat{\theta}_{l_1}) \right\} \\ & + \inf_{b \in (\bar{Y}_{k_j}, \bar{Y}_{k_j+1})} \left\{ \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - b)^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\hat{\theta}_{l_2} - b) \right\} \end{aligned} \tag{2.52}$$

We can observe that (2.52) is the sum of two copies of (2.46) in case (A1). Hence, by following the same arguments as before, we see that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

It therefore remains for us to obtain the result in the case that $b^* - a^* < \gamma\lambda$, and we denote this case (B2). Using that the separation (2.43) $\geq 3\gamma\lambda + 2\zeta$, it is straightforward to see that one of $(\bar{Y}_{k_j+1} - b^*)$ and $(a^* - \bar{Y}_{k_j})$ must be at least $\gamma\lambda + \zeta$. By the symmetry of the problem, it is sufficient for us to consider the case where $\bar{Y}_{k_j+1} - b^* \geq \gamma\lambda + \zeta$. In this case,

we can obtain from (2.50) that

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{(\tilde{a},\tilde{b})\in\mathcal{B}} \Bigg\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{b})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\tilde{b} - \tilde{a}) + \rho(\tilde{a} - \hat{\theta}_{l_1}) \Bigg\}, \tag{2.53}$$

where $\mathcal{B} = \Big\{(\tilde{a},\tilde{b})\colon \hat{\theta}_{l_1} \leq \tilde{a} \leq \tilde{b} \leq \bar{Y}_{k_j+1} - \gamma\lambda - \zeta,\ \tilde{b} - \tilde{a} < \gamma\lambda\Big\}$. From this, we can extract the terms dependent on $\tilde{b}$ to obtain

$$\tilde{b}^* = \operatorname*{arg\,min}_{\tilde{a}^*\leq\tilde{b}<\tilde{a}^*+\gamma\lambda} \left\{ \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{b})^2 + \rho(\tilde{b} - \tilde{a}^*) \right\}. \tag{2.54}$$

This objective is piecewise quadratic (and continuously differentiable), with two pieces; $[\tilde{a}^*, \tilde{a}^* + \gamma\lambda)$ and $[\tilde{a}^* + \gamma\lambda, \hat{\theta}_{l_2}]$. Over the second region, the objective is a convex quadratic with minimum at $\bar{Y}_{k_j+1} \in [\tilde{a}^* + \gamma\lambda, \hat{\theta}_{l_2}]$. By following the same argument as for (2.48) in case (A2), we see that $\tilde{b}^* = \tilde{a}^*$.

With this knowledge, we can further simplify (2.53) to obtain

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \geq \min_{\hat{\theta}_{l_1}\leq\tilde{a}\leq\bar{Y}_{k_j+1}-\gamma\lambda-\zeta} \Bigg\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 + \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 - \frac{\tilde{w}_{k_j+1}}{2}(\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}) \Bigg\}.$$

Since $\bar{Y}_{k_j+1} - \tilde{a}^* > \zeta$, we can see that $(\bar{Y}_{k_j+1} - \tilde{a}^*)^2 - (\bar{Y}_{k_j+1} - \hat{\theta}_{l_2})^2 > 0$. Thus, it suffices for us to show that

$$\min_{\hat{\theta}_{l_1}\leq\tilde{a}\leq\bar{Y}_{k_j+1}-\gamma\lambda-\zeta} \left\{ \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \tilde{a})^2 - \frac{\tilde{w}_{k_j}}{2}(\bar{Y}_{k_j} - \hat{\theta}_{l_1})^2 + \rho(\tilde{a} - \hat{\theta}_{l_1}) \right\} \geq 0.$$

This objective is exactly as in (2.47) in case (A1), minimised over a smaller feasible set. Hence, it follows immediately that this holds and we can conclude that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$.

We now have for all cases that $Q(\hat{\boldsymbol{\theta}}) > Q(\tilde{\boldsymbol{\theta}})$, which contradicts the optimality of $\hat{\theta}$. Thus, we can conclude that for $j = 1, \ldots, s$, $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j}$ and $\hat{\theta}_{k_{j-1}+1} \geq \bar{Y}_{k_{j-1}+1}$. □

Lemma 2.9. *Consider the setup of Lemma 2.8. For each $j = 1, \ldots, s$, there exists k_j^* in $\{k_{j-1} + 1, \ldots, k_j\}$ such that $\hat{\theta}_{k_j^*} \in [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$.*

Proof. We first show that if $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$, then for any k with $k_{j-1} + 1 \leq k \leq k_j$, if $\hat{\theta}_k > \bar{Y}_{k_j}$ then $\hat{\theta}_k = \hat{\theta}_{k_j}$.

We prove the first case since the proof for the second is identical. Suppose that this does not hold, i.e. $\hat{\theta}_{k_j} > \bar{Y}_{k_j}$ and there exists some (minimal) k in $\{k_{j-1}+1, \ldots, k_j - 1\}$ with $\bar{Y}_{k_j} < \hat{\theta}_k < \hat{\theta}_{k_j}$. Then we construct $\check{\boldsymbol{\theta}}$ by

$$\check{\theta}_l = \begin{cases} \hat{\theta}_k & \text{for } l = k, k+1, \ldots, k_j \\ \hat{\theta}_l & \text{otherwise.} \end{cases} \tag{2.55}$$

We observe that the penalty contribution from $\check{\boldsymbol{\theta}}$ is no more than that of $\hat{\boldsymbol{\theta}}$ and that the quadratic loss for $\check{\boldsymbol{\theta}}$ will be strictly less than that of $\hat{\boldsymbol{\theta}}$. This gives us that $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\boldsymbol{\theta}}$.

Similarly, if $\hat{\theta}_{k_{j-1}+1} < \bar{Y}_{k_{j-1}+1}$ then the corresponding statement that for any k with $k_{j-1}+1 \leq k_j$, if $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$ then $\hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}$.

We now establish a simple preliminary result. Suppose that for some j in $\{1, \ldots, s\}$ there exists k in $\{k_{j-1}+1, \ldots, k_j\}$ with $\hat{\theta}_k \notin [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$, such that $\sum_{\{l:\, \hat{\theta}_l = \hat{\theta}_k\}} w_l \geq \eta/2s$. We claim that if $\hat{\theta}_k > \bar{Y}_{k_j}$ then $\hat{\theta}_k \leq \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$. Similarly, if $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$ then $\hat{\theta}_k \geq \bar{Y}_{k_{j-1}+1} - (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$.

To prove the claim, we consider the case $\hat{\theta}_k > \bar{Y}_{k_j}$ (the other is identical). By the first observation, if $\hat{\theta}_l > \bar{Y}_{k_j}$ for l in $\{k_{j-1}+1, \ldots k_j\}$ then $\hat{\theta}_l = \hat{\theta}_k$. Now, for contradiction, suppose $\hat{\theta}_k > \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$ and let this k be minimal. Then we can construct $\check{\boldsymbol{\theta}}$ by

$$\check{\theta}_l = \begin{cases} \sum_{l=k}^{k_j} w_l \bar{Y}_l / \sum_{l=k}^{k_j} w_l & \text{for } l = k, \ldots, k_j \\ \hat{\theta}_l & \text{otherwise.} \end{cases}$$

By appealing to the optimality of $\hat{\boldsymbol{\theta}}$, we can easily observe that $\hat{\theta}_{k-1} \leq \bar{Y}_{k-1}$ and therefore that the ordering of the entries of $\check{\boldsymbol{\theta}}$ matches that of $\hat{\boldsymbol{\theta}}$. Here, we use that $(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) \geq \gamma\lambda$.

We can now see that the loss term in $Q(\check{\boldsymbol{\theta}})$ is less than in $Q(\hat{\boldsymbol{\theta}})$, with a difference of more than $(\eta/4s)(\sqrt{2s/\eta}\sqrt{\gamma}\lambda)^2 = \gamma\lambda^2/2$, which outweighs the possible increase in the penalty contribution. This gives us that $Q(\check{\boldsymbol{\theta}}) < Q(\hat{\boldsymbol{\theta}})$, contradicting the optimality of $\hat{\boldsymbol{\theta}}$.

We now return to the proof of the main result. Suppose, for contradiction, that there exists some $j \in \{1, \ldots, s\}$ such that $\hat{\theta}_k \notin [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$ for all $k = k_{j-1}+1, \ldots, k_j$ and let this j be minimal. By the first observation, we know that entries of $\hat{\boldsymbol{\theta}}$ corresponding to level j can take one of at most two distinct values. That is, for $k \in \{k_{j-1}+1, \ldots, k_j\}$, if we have $\hat{\theta}_k < \bar{Y}_{k_{j-1}+1}$, then it follows that $\hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}$. Similarly, if $\hat{\theta}_k > \bar{Y}_{k_j}$, then $\hat{\theta}_k = \hat{\theta}_{k_j}$.

By the assumption $w^0_{\min} \geq \eta/s$, we have that either

$$\sum_{k:\, \hat{\theta}_k = \hat{\theta}_{k_{j-1}+1}} w_k \geq \frac{\eta}{2s} \quad \text{or} \quad \sum_{k:\, \hat{\theta}_k = \hat{\theta}_{k_j}} w_k \geq \frac{\eta}{2s}.$$

We will without loss of generality take the second statement to be true (the proof for the first case follows identically). Let k' denote the minimal element in $\{k_{j-1}+1,\dots,k_j\}$ such that $\hat{\theta}_{k'} = \hat{\theta}_{k_j}$. From the preliminary result established earlier, $\hat{\theta}_{k_j} \leq \bar{Y}_{k_j} + (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda)$. By appealing to the optimality of $\hat{\boldsymbol{\theta}}$, we see that $\hat{\theta}_{k_j+1} < \hat{\theta}_{k_j} + \gamma\lambda$ (otherwise, we could take $\hat{\theta}_{k_j}$ to be $\bar{Y}_{k_j}$ and strictly reduce the value of the objective).

Now, we will use that the separation is at least $2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda$. By our earlier observation (2.55), it is clear that any $l \in \{k_j+1,\dots,k_{j+1}\}$ with $\hat{\theta}_l < \bar{Y}_{k_j+1}$ has $\hat{\theta}_l = \hat{\theta}_{k_j+1}$. Note that since $\hat{\theta}_{k_j+1} - \bar{Y}_{k_j} < (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda$, it follows that $\bar{Y}_{k_j+1} - \hat{\theta}_{k_j+1} > (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta$ and therefore that $\sum_{\{k:\, \hat{\theta}_k = \hat{\theta}_{k_j+1}\}} w_k < \eta/2s$ by the preliminary result. Since $w^0_{\min} \geq \eta/s$ and separation (2.43) $\geq 2(\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \gamma\lambda + \zeta$, we can define $l' \in \{k_j+1,\dots,k_{j+1}\}$ minimal such that $\hat{\theta}_{l'} \geq \bar{Y}_{k_j+1}$.

Now, in order to contradict the optimality of $\hat{\boldsymbol{\theta}}$ we construct a new feasible point $\tilde{\boldsymbol{\theta}}$ by setting

$$\tilde{\theta}_l = \begin{cases} \bar{Y}_{k_j} & \text{for } l = k',\dots,k_j \\ \hat{\theta}_{l'} & \text{for } l = k_j+1,\dots,l'-1 \\ \hat{\theta}_l & \text{otherwise.} \end{cases}$$

It follows that for $l = k_j+1,\dots,l'-1$ we have

$$\begin{aligned} |\hat{\theta}_l - \bar{Y}_l| &> (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta \\ |\tilde{\theta}_l - \bar{Y}_l| &\leq (\sqrt{2s/\eta}\sqrt{\gamma}\lambda \vee \gamma\lambda) + \zeta. \end{aligned}$$

It is also straightforward to see that $|\hat{\theta}_{k_j} - \bar{Y}_l| \geq |\bar{Y}_{k_j} - \bar{Y}_l|$ for $l = k',\dots,k_j$. If follows that the loss contribution in $Q(\tilde{\boldsymbol{\theta}})$ is strictly less than that in $Q(\hat{\boldsymbol{\theta}})$. Hence, using $\hat{\theta}_{l'} - \hat{\theta}_{k_j} > \gamma\lambda$, we obtain

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) >& \rho(\hat{\theta}_{l'} - \hat{\theta}_{k_j+1}) + \rho(\hat{\theta}_{k_j+1} - \hat{\theta}_{k_j}) + \rho(\hat{\theta}_{k_j} - \hat{\theta}_{k'-1}) \\ &- \frac{1}{2}\gamma\lambda^2 - \rho(\bar{Y}_{k_j} - \hat{\theta}_{k'-1}) \\ \geq& 0, \end{aligned}$$

contradicting the optimality of $\hat{\boldsymbol{\theta}}$. We conclude that for $j = 1,\dots,s$, there exists k_j^* in $\{k_{j-1}+1,\dots,k_j\}$ such that $\hat{\theta}_{k_j^*} \in [\bar{Y}_{k_{j-1}+1}, \bar{Y}_{k_j}]$. □

Lemma 2.10. *Consider the univariate objective* (2.11)*, relaxing the normalisation constraint to* $\breve{w} = \sum_k w_k \leq 1$*. Suppose that* $w^T\bar{\boldsymbol{Y}} = 0$*, and that* $\|\bar{\boldsymbol{Y}}\|_\infty < (2 \wedge \sqrt{\gamma\breve{w}})\,\lambda/\breve{w}$*. Then* $\hat{\boldsymbol{\theta}} = 0$*.*

Proof. Let $P_w = I - \mathbf{1}w^T/\breve{w}$ and $D_w \in \mathbb{R}^{K\times K}$ be the diagonal matrix with entries $D_{kk}\sqrt{w_k}$. First note that

$$\begin{aligned} Q(\boldsymbol{\theta}) - Q(P_w\boldsymbol{\theta}) &= \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \theta_k)^2 - \frac{1}{2}\sum_{k=1}^{K} w_k(\bar{Y}_k - \theta_k + w^T\boldsymbol{\theta})^2 \\ &= -\frac{1}{2}\sum_{k=1}^{K} w_k(w^T\boldsymbol{\theta})(2\bar{Y}_k - 2\theta_k + w^T\boldsymbol{\theta}) \\ &= \left(1 - \frac{1}{2}\breve{w}\right)(w^T\boldsymbol{\theta})^2 \ \geq 0. \end{aligned}$$

Thus for all $\boldsymbol{\theta} \in \mathbb{R}^K$, we have

$$\begin{aligned} Q(\boldsymbol{\theta}) - Q(0) &\geq \frac{1}{2}\|D_wP_w(\bar{\boldsymbol{Y}} - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\bar{\boldsymbol{Y}}\|_2^2 + \sum_{k=1}^{K-1}\rho(\theta_{(k+1)} - \theta_{(k)}) \\ &\geq \frac{1}{2}\|D_wP_w(\bar{\boldsymbol{Y}} - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\bar{\boldsymbol{Y}}\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)}) \\ &\geq \min_{\xi\in[-\tau,\tau]^K} F(\boldsymbol{\theta}, \xi, w) \end{aligned}$$

where

$$F(\boldsymbol{\theta}, \xi, w) = \frac{1}{2}\|D_wP_w(\xi - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_wP_w\xi\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)}).$$

Consider minimising F over $\mathbb{R}^K \times [-\tau,\tau]^K \times S$, where $S \subseteq \mathbb{R}^K$ is the unit simplex scaled by $\breve{w}$. We aim to show this minimum is 0. As with the first claim in the proof of Lemma 2.8, it is straightforward to see that for any feasible $(\boldsymbol{\theta}, \xi, w)$, there exists $\boldsymbol{\theta}'$ with $\|\boldsymbol{\theta}'\|_\infty \leq \|\xi\|_\infty$ and $F(\boldsymbol{\theta}', \xi, w) \leq F(\boldsymbol{\theta}, \xi, w)$. Hence,

$$\inf_{(\boldsymbol{\theta},\xi,w)\in\mathbb{R}^K\times[-\tau,\tau]^K\times S} F(\boldsymbol{\theta}, \xi, w) = \inf_{(\boldsymbol{\theta},\xi,w)\in[-\tau,\tau]^K\times[-\tau,\tau]^K\times S} F(\boldsymbol{\theta}, \xi, w).$$

As on the RHS we are minimising a continuous function over a compact set, we know a minimiser must exist. Let $(\tilde{\boldsymbol{\theta}}, \tilde{\xi}, \tilde{w})$ be a minimiser (to be specified later). Observe that

$$\|D_{\tilde{w}}P_{\tilde{w}}(\xi - \boldsymbol{\theta})\|_2^2 - \|D_{\tilde{w}}P_{\tilde{w}}\xi\|_2^2 = -2\xi^TP_{\tilde{w}}^TD_{\tilde{w}}^2P_{\tilde{w}}\boldsymbol{\theta} + \theta^TP_{\tilde{w}}^TD_{\tilde{w}}^2P_{\tilde{w}}\boldsymbol{\theta}$$

is linear as a function of ξ. Hence it is minimised over the set

$$\{\xi : \|\xi\|_\infty \leq \tau\} = \mathrm{conv}(\{-\tau, \tau\}^K)$$

at some point in $\{-\tau,\tau\}^K$. Here $\mathrm{conv}(\cdot)$ denotes the convex hull operation. We thus have

$$Q(\boldsymbol{\theta}) - Q(0) \geq \min_{\xi\in\{-\tau,\tau\}^K}\left\{\frac{1}{2}\|D_{\tilde{w}}P_{\tilde{w}}(\xi - \boldsymbol{\theta})\|_2^2 - \frac{1}{2}\|D_{\tilde{w}}P_{\tilde{w}}\xi\|_2^2 + \rho(\theta_{(K)} - \theta_{(1)})\right\}.$$

Let us take $(\tilde{\boldsymbol{\theta}}, \tilde{\xi}) \in \mathbb{R}^K \times \{-\tau, \tau\}^K$ to be a minimiser of the RHS.

Note that if we have $\tilde{\xi}_j = \tilde{\xi}_k$ then we may take $\tilde{\theta}_j = \tilde{\theta}_k$. Indeed, we may construct $\check{\boldsymbol{\theta}} \in \mathbb{R}^K$ by setting

$$\check{\theta}_l = \begin{cases} \arg\min_{b \in \{\tilde{\theta}_j, \tilde{\theta}_k\}} (\tilde{\xi}_j - b)^2 & \text{for } l = j, k \\ \tilde{\theta}_l & \text{otherwise.} \end{cases}$$

Since the penalty contribution from $\check{\boldsymbol{\theta}}$ is not greater than that of $\tilde{\boldsymbol{\theta}}$, it follows that $Q(\check{\boldsymbol{\theta}}) \leq Q(\tilde{\boldsymbol{\theta}})$. Thus, we can assume that entries of $\tilde{\boldsymbol{\theta}}$ can take one of only two distinct values.

Next we write $\tilde{\alpha} = \sum_{k:\tilde{\xi}_k = -\tau} \tilde{w}_k$ and observe that $\tilde{w}^T \tilde{\xi} = (\check{w} - 2\tilde{\alpha})\tau$. Let us set $s = \min_k \tilde{\theta}_k$ and $x = \max_k \tilde{\theta}_k - \min_k \tilde{\theta}_k$. Then we have

$$\begin{aligned} F(\tilde{\boldsymbol{\theta}}, \tilde{\xi}, \tilde{w}) =& \frac{1}{2}\tilde{\alpha}\{(2\tilde{\alpha} - 1 - \check{w})\tau - s\}^2 + \frac{1}{2}(\check{w} - \tilde{\alpha})((2\tilde{\alpha} + 1 - \check{w})\tau - s - x)^2 \\ &+ \rho(x) - \frac{2}{\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})\tau^2 \\ =& \frac{1}{2\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})(2\tau - x)^2 + \rho(x) - \frac{2}{\check{w}}\tilde{\alpha}(\check{w} - \tilde{\alpha})\tau^2 \\ =& \frac{\check{w}}{8}(2\tau - x)^2 + \rho(x) - \frac{1}{2}\tau^2. \end{aligned} \tag{2.56}$$

In the second line above, we have solved for s to find that

$$s = \frac{1}{\check{w}}\{\tau(1 - \check{w})(\check{w} - 2\tilde{\alpha}) + (\tilde{\alpha} - \check{w})x\}.$$

In the third line above, we have solved for $\tilde{\alpha}$ to obtain $\tilde{\alpha} = \check{w}/2$ and hence $\tilde{\alpha}(\check{w} - \tilde{\alpha})/\check{w} = \check{w}/4$. These follow from optimality of $\tilde{\boldsymbol{\theta}}$ and $\tilde{w}$ respectively. The result follows from applying Lemma 2.7, setting $\kappa = \check{w}/4$. □

2.10.2 Proof of Theorem 2.6

We begin by defining P^0 to be the orthogonal projection onto the linear space

$$V_0 = \left\{ \mu + \sum_{j=1}^{j} \sum_{k=1}^{K_j} \mathbb{1}_{\{X_{ij}=k\}} \theta_{jk} : (\mu, \theta) \in \mathbb{R} \times \Theta_0 \right\}.$$

The residuals from the oracle least squares fit are $(I - P^0)\boldsymbol{Y} = (I - P^0)\varepsilon$. The partial residuals $R^{(j)}$ as defined in (2.18) for the j^{th} variable are therefore

$$R_i^{(j)} = \sum_{k=1}^{K_j} \mathbb{1}_{\{X_{ij}=k\}} \hat{\theta}^0_{jk} + \left[(I - P^0)\varepsilon\right]_i. \tag{2.57}$$

For $j = 1, \ldots, p$, we define $\bar{R}_k^{(j)} = \sum_{i=1}^n \mathbb{1}_{\{X_{ij}=k\}} R_i^{(j)}/n_{jk}$ for $k = 1, \ldots, K_j$, reordering the labels such that $\bar{R}_1^{(j)} \leq \cdots \leq \bar{R}_{K_j}^{(j)}$. We then aim to apply the arguments of Theorem 2.5

to $\hat{\boldsymbol{\theta}}_j$ defined by

$$\hat{\boldsymbol{\theta}}_j \in \underset{\boldsymbol{\theta}_j \in \Theta_j}{\arg\min} \left\{ \frac{1}{2} \sum_{k=1}^{K_j} w_{jk} \left(\bar{R}_k^{(j)} - \theta_{jk} \right)^2 + \sum_{k=1}^{K_j - 1} \rho(\theta_{jk+1} - \theta_{jk}) \right\}. \tag{2.58}$$

In order to do this, we define the events (for some τ_j to be determined later):

$$\Lambda_j^{(1)} = \left\{ |\hat{\theta}^0_{jk_l} - \theta^0_{jk_l}| \leq \tau_j : l = 1, \ldots, s_j \right\} \tag{2.59}$$

$$\Lambda_{jk}^{(2)} = \left\{ \left| \frac{1}{n_{jk}} \sum_{i=1}^{n} \mathbb{1}_{\{X_{ij}=k\}} ((I - P^0)\varepsilon)_i \right| < \frac{1}{2} \sqrt{\eta \gamma_{*j} s_j} \lambda_j \right\}. \tag{2.60}$$

On the intersection of events $\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)}$, we have that $|\bar{R}_k^{(j)} - \hat{\theta}^0_{jk}| < \sqrt{\eta \gamma_{*j} s_j} \lambda_j / 2$. By following an identical approach to that involved in computing (2.35), we have that

$$\mathbb{P}\left(\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)} \right) \geq 1 - 2 \exp\left(- \frac{n w_{j,\min} \eta \gamma_{*j} s_j \lambda_j^2}{8\sigma^2} + \log(K_j) \right),$$

where we recall that $w_{jk} = n_{jk}/n$.

We now turn our attention to the event $\Lambda_j^{(1)}$. Note that if $s_j = 1$, then this is immediately satisfied since $\hat{\boldsymbol{\theta}}_j^0 = \boldsymbol{\theta}_j^0 = 0$. If $s_j > 1$, we use that the oracle least squares estimate $\hat{\boldsymbol{\theta}}^0 = A\boldsymbol{Y}$ is a linear transformation A of the responses $(Y_i)_{i=1}^n$. For each $i = 1, \ldots, n$, Y_i has an independent (non-central) sub-Gaussian distribution with parameter σ. Therefore for each $k = 1, \ldots, K_j$, $\hat{\theta}^0_{jk} - \theta^0_{jk}$ also has a sub-Gaussian distribution, with parameter at most $\sigma c_{\min}^{-1/2}$ (recalling that $c_{\min} = (\max_l (AA^T)_{ll})^{-1}$). This enables us to show that

$$\mathbb{P}\left(\Lambda_j^{(1)} \right) \geq 1 - 2 \exp\left(- \frac{c_{\min} \tau_j^2}{2\sigma^2} + \log(s_j) \right).$$

We can now set $\tau_j = \sqrt{\eta \gamma_{*j} s_j} \lambda_j / 2$. From (2.26) and the triangle inequality, on the event $\Lambda_j^{(1)}$ we have that

$$\begin{aligned} \Delta(\hat{\boldsymbol{\theta}}_j^0) &\geq \Delta(\boldsymbol{\theta}_j^0) - \sqrt{\eta \gamma_{*j} s_j} \lambda_j \\ &\geq 3 \left(1 + \frac{\sqrt{2}}{\eta} \right) \sqrt{\gamma_j \gamma_j^*} \lambda_j. \end{aligned}$$

Thus, on the intersection of events $\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j} \Lambda_{jk}^{(2)} \right)$, we can proceed as in the proof of Theorem 2.5 from (2.38), to conclude that $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}_j^0$.

It immediately follows that on the intersection of events $\cap_{j=1}^{p}\left(\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j}\Lambda_{jk}^{(2)}\right)\right)$, we have $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^0$. By a union bound, this occurs with probability at least

$$\begin{aligned}\mathbb{P}\left(\cap_{j=1}^{p}\left(\Lambda_j^{(1)} \cap \left(\cap_{k=1}^{K_j}\Lambda_{jk}^{(2)}\right)\right)\right) &\geq 1 - 2\sum_{j=1}^{p}\left[\exp\left(-\frac{n_{j,\min}\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(K_j)\right)\right. \\ &\qquad \left. + \exp\left(-\frac{c_{\min}\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(s_j)\right)\right] \\ &\geq 1 - 4\sum_{j=1}^{p}\exp\left(-\frac{(n_{j,\min} \wedge c_{\min})\eta\gamma_{*j}s_j\lambda_j^2}{8\sigma^2} + \log(K_j)\right),\end{aligned}$$

where in the final line we use $s_j \leq K_j$. □

2.11 Supplementary material for Section 2.6

2.11.1 Details of methods

Tree-based methods

We used the implementation of the random forest procedure (Breiman, 2001) in the R package `randomForest` (Liaw and Wiener, 2002) with default settings. CART (Breiman et al., 1984) was implemented in the R package `rpart` (Therneau and Atkinson, 2019), with pruning according to the 1-SE rule (as described in the package documentation).

CAS-ANOVA

The CAS-ANOVA estimator $\hat{\boldsymbol{\theta}}^{\text{cas}}$ optimises over $(\mu, \boldsymbol{\theta})$ a sum of a squared loss term (2.3) and an all-pairs penalty term (2.4). In particular, Bondell and Reich (2009) consider two regimes of weight vectors w. The first is not data-dependent and sets $w_{j,k_1k_2} = (K_j+1)^{-1}\sqrt{n_{jk_1} + n_{jk_2}}$. The second, 'adaptive CAS-ANOVA', uses the ordinary least squares estimate for $\boldsymbol{\theta}$ to scale the weights. Here, $w_{j,k_1k_2} = (K_j+1)^{-1}\sqrt{n_{jk_1} + n_{jk_2}}|\hat{\theta}^{\text{OLS}}_{jk_1} - \hat{\theta}^{\text{OLS}}_{jk_2}|^{-1}$.

Here we introduce a new variant of adaptive CAS-ANOVA, following ideas in Zou (2006) for a 2-stage adaptive Lasso procedure. Instead of using the ordinary least squares estimate $\hat{\boldsymbol{\theta}}^{\text{OLS}}$ in the above expression, an initial (standard) CAS-ANOVA estimate is used to scale the weights, with λ selected for the initial estimate by 5-fold cross-validation. In simulations, this outperformed the adaptive CAS-ANOVA estimate using ordinary least squares initial estimates so in the interests of time and computational resources this was omitted from the simulation study. Henceforth adaptive CAS-ANOVA will refer to this 2-stage procedure.

The authors describe the optimisation of $\hat{\boldsymbol{\theta}}^{\text{cas}}$ as a quadratic programming problem, which was solved using the R package `rosqp` (Anderson, 2018). Here we used our own implementation of the quadratic programming approach described by the authors. We found

it considerably faster than the code available from the authors' website, and uses ADMM-based optimisation (Boyd et al., 2011) tools not available at the time of its publication. We also found, as discussed in Section 5.1 of Maj-Kańska et al. (2015), that we could not achieve the best results using the publicly available code. Lastly, using our own implementation allowed us to explore a modification of CAS-ANOVA using the more modern approach of adaptive weights via a 2-stage procedure (Bühlmann and van de Geer, 2011) to compare SCOPE to a wider class of all-pairs penalty procedures.

For large categorical variables, solutions are slow to compute and consume large amounts of memory. In the case of binary response, CAS-ANOVA models were fitted iterating a locally quadratic approximation to the loss function.

DMR

The DMR algorithm (Maj-Kańska et al., 2015) is implemented in the R package `DMRnet` (Prochenka-Sołtys and Pokarowski, 2018). The degrees of freedom in the model is decided by 5-fold cross-validation. It is based on pruning variables using the Group Lasso (Yuan and Lin, 2006) to obtain at a low-dimensional model, then performing backwards selection based on ranking t-statistics for hypotheses corresponding to each fusion between levels in categorical variables.

The cross-validation routine appeared to error when all levels of all categorical variables were not present in one of the folds. In Section 2.6.2, cross-validation was therefore not possible so model selection was performed based on Generalized Information Criterion (GIC) (Zheng and Loh, 1995). In all other examples, models were selected via 5-fold cross-validation.

Bayesian effect fusion

In Section 2.6.1 we include Bayesian effect fusion (Pauger and Wagner, 2019), implemented in the R package `effectFusion` (Pauger et al., 2019). Coefficients within each categorical variable were modelled with a sparse Gaussian mixture model. The posterior mean was estimated with 1000 samples.

Lasso

In Section 2.6.1 we also include Lasso (Tibshirani, 1996) fits, to serve as a reference point. Of course, this is unsuitable for models where levels in categorical variables should be clustered together, but the advanced development of the well-known R package `glmnet` (Friedman et al., 2010) nevertheless sees its use in practice.

In order to make the fit symmetric across the categories within each variable, models were fitted with an unpenalised intercept and featuring dummy variables for all of the categories within each variable. This is instead of the corner-point dummy variable encoding

of factor variables that is commonly used when fitting linear models. Models are fitted and cross-validated with `cv.glmnet` using the default settings.

SCOPE

For SCOPE, we have provided the R package `CatReg` (Stokell et al., 2021a). The univariate update step (see Section 2.3.1) is implemented in C++ using Rcpp (Eddelbuettel and François, 2011), with models fitted using a wrapper in R. For the binary response case, the outer loop to iterate the local quadratic approximations in the proximal Newton algorithm are done within R. In the future, performance could be improved by iterating the univariate update step (and the local quadratic approximations, as in Sections 2.6.2 and 2.6.2) within some lower-level language. In higher-dimensional experiments, SCOPE was slowed by cycling through all the variables; an active-set approach to this could make it faster still.

2.11.2 Further details of numerical experiments

For the experiments in Section 2.6.1, we define the signal-to-noise ratio (SNR) as σ_S/σ, where σ_S is the standard deviation of the signal $Y - \varepsilon$, and σ is the standard deviation of the noise ε.

Low-dimensional simulations

In Table 2.7 we include details of computation time and dimension of the fitted models. Figure 2.9 visualises the results also summarised in Table 2.1 in Section 2.6.1.

	Mean fitted model dimension				
σ^2:	1	6.25	25	100	Mean computation time (s)
SCOPE-8	7.2	8.5	4.7	4.3	16
SCOPE-32	9.6	12.6	13.2	9.8	48
SCOPE-CV	7.9	10.3	16.8	10.9	68
Oracle least squares	7.0	7.0	7.0	7.0	0.00
Linear regression	231.0	231.0	231.0	231.0	0.01
CAS-ANOVA	35.2	70.0	74.3	52.4	4679
Adaptive CAS-ANOVA	13.4	31.3	36.9	32.5	9659
DMR	7.0	7.2	5.3	2.7	21
BEF	7.3	6.3	4.1	2.0	975
CART					0.01
RF					0.66

Table 2.7 Mean fitted model dimension and computation time for the various methods.

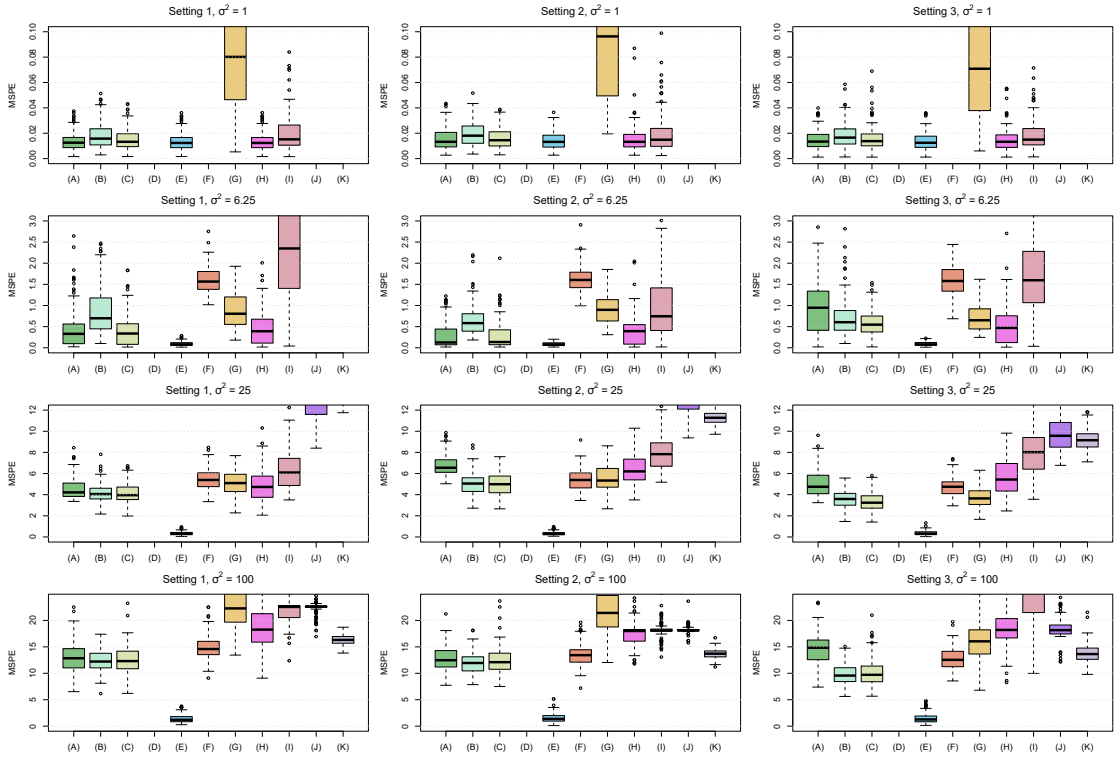


Figure 2.9 Prediction performance of various methods: (A) SCOPE-8; (B) SCOPE-32; (C) SCOPE-CV; (D) Linear regression; (E) Oracle least squares; (F) CAS-ANOVA; (G) Adaptive CAS-ANOVA; (H) DMR; (I) BEF; (J) CART; (K) RF. Note that some 'boxes' are not visible in some of the plots; this is due to the MSPE in the tests being beyond the range of the plot.

High-dimensional simulations

Here we include additional results relating to the high-dimensional experiments. Figure 2.10 visualises the results in Table 2.2 in Section 2.6.1.

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	224	322	348	76	234	518	209	175
SCOPE-32	134	341	502	51	283	650	113	161
SCOPE-CV	951	1739	2450	332	1516	2892	767	902
DMR	26	38	39	26	30	36	30	29
CART	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.1
RF	5.7	5.7	5.9	2.7	5.8	5.8	5.9	5.8
Lasso	1.5	1.5	1.6	1.2	1.4	1.5	1.5	1.5

Table 2.8 Mean computation time (s)

Setting:	1	2	3	4	5	6	7	8
SCOPE-8	6.9	9.4	9.8	6.9	21.3	27.1	9.3	7.2
SCOPE-32	20.7	37.5	38.0	19.9	75.8	26.1	32.9	31.3
SCOPE-CV	21.4	40.4	40.8	19.5	103.7	26.2	36.6	17.9
DMR	1.9	4.9	4.7	3.4	3.7	22.8	2.3	7.5
Lasso	15.7	167.1	152.0	32.7	143.7	469.7	35.8	82.8

Table 2.9 Mean fitted model dimension

Setting	γ:	4	8	16	32	64
1		0.028	0.290	0.196	0.138	**0.348**
2		0.002	0.016	0.234	0.298	**0.450**
3		0.006	0.012	0.286	0.248	**0.448**
4		0.030	**0.356**	0.244	0.100	0.270
5		0.000	0.000	0.026	0.070	**0.904**
6		0.000	0.000	0.464	**0.534**	0.002
7		0.006	0.092	0.234	0.144	**0.524**
8		0.264	**0.446**	0.102	0.018	0.170

Table 2.10 Proposition of times each γ was selected by cross-validation.

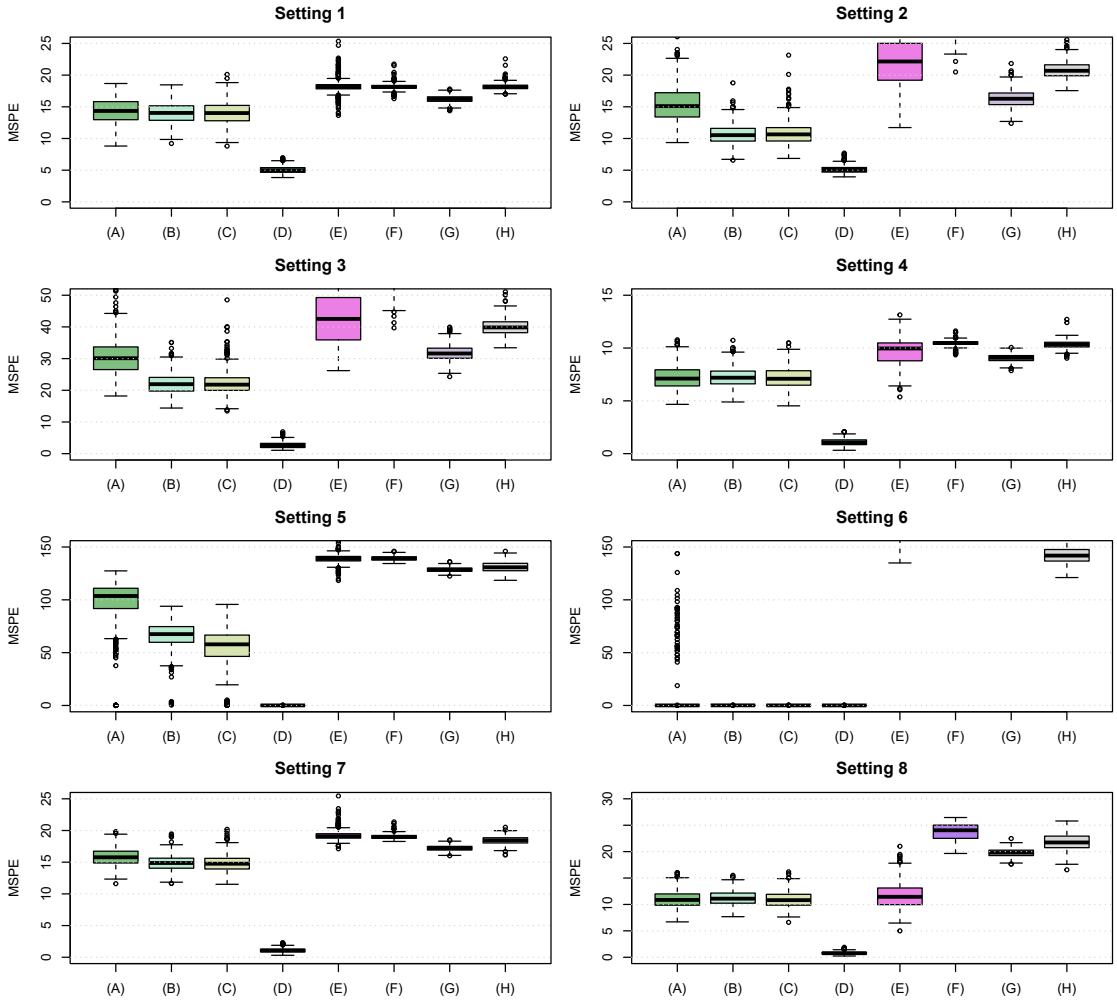


Figure 2.10 Prediction performance of various methods: (A) SCOPE-8; (B) SCOPE-32; (C) SCOPE-CV; (D) Oracle least squares; (E) DMR; (F) CART; (G) RF; (H) Lasso. Note that some 'boxes' are not visible in some of the plots; this is due to the MSPE in the tests being beyond the range of the plot.

Chapter 3

High-dimensional regression with potential prior information

3.1 Introduction

An important goal of the Lasso (Tibshirani, 1996), which given a response $\boldsymbol{Y} \in \mathbb{R}^n$ and matrix of predictors $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ solves

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \hat{\mu}\mathbf{1} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{3.1}$$

and the many related penalised regression procedures overviewed in Section 1.1, is to determine which variables are relevant. In many settings, some vague prior information on the relative importance of the variables may be available. One instance of this concerns the scaling of variables so they have the same empirical variance before performing the Lasso optimisation (3.1). Unlike the setting investigated in Chapter 2, where the covariates are categorical and thus with no scale information to consider when fitting a model such as SCOPE, the scaling of coefficients of continuous predictors depends on that of its column in the design matrix. This practice of rescaling all of the columns to have unit norm is very common and is carried out by default in many software packages including the highly popular `glmnet` (Friedman et al., 2010). The rationale for this is to ensure that all of the coefficients are treated in a balanced way; otherwise, coefficients corresponding to variables with large empirical variance will effectively experience very little shrinkage whereas those corresponding to variables with small variances will be penalised heavily.

On the other hand, any information that may be encoded in the scale of the columns of $\boldsymbol{X}$ is lost. For example, in a setting where one may expect measurement error to be distributed evenly over the variables, it is reasonable to suspect that variables with larger observed variance will be less corrupted by the error and hence contain more underlying signal. Indeed,

it is also common to remove columns with the smallest variance as a pre-processing step[1]. This is however a somewhat crude way of using this potential information, and invites the possibility of the scale information being used more systematically.

An a priori ordering on variables could arise for many other reasons. For example, in a time-lagged regression, where an outcome at each time point is predicted using variables from a fixed number of previous time points, it is natural to assume that the importance of variables decreases with the lag, implying an ordering. There may also be periodicity or seasonality considerations that can be incorporated into the ordering. Alternatively, in a regression model where there is a spatial structure over the variables, it is reasonable to suspect that measurements from further away may be less informative. Based on subject knowledge and prior experimental results, some variables may already be known to be important, in which case a (possibly partial) ordering over the variables will be readily available. We will focus on the general question of how such an ordering can be used to improve models.

One approach to incorporating this information involves modifying the penalty term allowing individual tuning parameters λ_j to be applied to each variable, $\sum_{j=1}^{p} \lambda_j |\beta_j|$, the idea being to place a smaller penalty on those coefficients believed to be more important. Manual reweighting of the penalty terms is computationally very attractive since it is no more difficult to compute than the Lasso (3.1), for which there exist very fast and reliable algorithms that include this functionality (Friedman et al., 2010). Nardi and Rinaldo (2011) study this approach in the autoregressive model setting. However, if the ordering is not helpful (perhaps even actively unhelpful), this approach runs into problems. If some important variables are given a larger penalty, the prediction and estimation performance of the model will become accordingly worse, and the fitted model can end up very different from that fitted with the usual Lasso penalty. The relatively large cost associated with being wrong makes this approach less attractive in practice.

Motivated primarily by the time-lagged regression example, Tibshirani and Suo (2016) propose fitting the Lasso with a monotonicity constraint imposed on the coefficients with respect to the natural ordering. However, their experimental results also suggest that the practitioner pays a large price for supplying a non-informative (e.g. randomised) ordering. Micchelli et al. (2010) consider a general norm-based penalty framework that includes, as an example, a penalty applied to groups (Yuan and Lin, 2006) which respect a specified ordering but are data-dependent. Such an approach, while less aggressive than a monotonicity constraint on the coefficients or a manual reweighting of the penalty as discussed above, still suffers in the event that the ordering is unhelpful.

In this work we propose to incorporate a potentially useful ordering over the variables in the following simple way, outlined in more detail in Section 3.2. We begin by fitting a model over the full set of variables, then proceed by fitting a sequence of submodels, each time

[1]https://www.kdnuggets.com/2015/05/7-methods-data-dimensionality-reduction.html

removing a subset consisting of the 'least important' variables according to our ordering. A model is then selected by validation using an appropriate prediction error metric. We show that for certain models of interest, including the Lasso (3.1), the computational price to pay for following our approach is small, enabling its use in many practical settings.

We show theoretically and empirically that if the ordering is informative, substantial improvements in terms of prediction error can be enjoyed, with performance close to that of the submodel selected with oracular knowledge of whichever happens to be best. Conversely, if the ordering is unhelpful, then compared to discarding it and fitting the full model, only a small statistical price is to be paid for following this approach.

3.1.1 Organisation of the chapter

In Section 3.2 we outline our method within a general regression setting, for Sections 3.2.2 and 3.2.3 to subsequently provide details of efficient computational strategies for Lasso and ridge regression models, both key models of interest. We include in Section 3.3 a result underpinning the strong performance of our approach in practice, showing that the prediction performance of the resulting model is only slightly worse than that enjoyed by strongest candidate. In Section 3.4 we outline an extension of our approach for fitting Lasso and ridge regression models based on potential prior information in the form of a nested tree structure, providing an interesting generalisation to new settings. Section 3.5 discusses the results from a range of experiments, including a simulation study and a number of real data examples.

3.2 Methodology

In this section we present our framework for systematically using potential prior ordering information when fitting high-dimensional regression models. While our method is applicable very generally, we will subsequently focus on the linear regression setting. For both Lasso and ridge regression models, we provide computational strategies that ensure our approach is practical and fast. Throughout this chapter, we will make use of the notation $[p] = \{1, \ldots, p\}$. A working version of our R software package is available online[2].

3.2.1 Using ordering information

Consider a general regression problem, with $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, of the form

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \ell(\boldsymbol{Y}, \boldsymbol{X}\beta) + \mathcal{P}(\beta) \right\}, \tag{3.2}$$

[2]https://github.com/bgs25/SubsetGridRegression

where ℓ is some loss function and $\mathcal{P}$ is a penalty term (such as $\lambda\|\beta\|_1$ in (3.1)). Suppose that we have an ordering $\pi \in \mathcal{S}_p$ (the symmetric group of order p) over the variables, such that we believe the signal variables, contained in a set $S \subseteq [p]$, appear early in this ordering.

In order to make use of this, we construct a grid of subsets of variables $[p] = S^1 \supseteq S^2 \supseteq \ldots \supseteq S^r$ by sequentially removing variables in the reverse order of π. For each subset S^l, we fit the model (3.2) restricted to just the variables in S^l. The candidate models are then compared by an appropriate prediction error metric on unseen test data, possibly within a cross-validation framework.

For a given grid size r, we select the grid such that the subsets are exponentially decreasing in size, i.e. $|S^l|/|S^{l-1}| \approx p^{1/(r-1)}$. This is motivated in part by the logarithmic cost associated with including extraneous variables in a Lasso regression. Compared to the smallest correct model possible when including variables sequentially according to π, $\min\{k\colon S \subseteq \{\pi(1), \ldots, \pi(k)\}\}$, if one can find $\arg\max_{l \in 1,\ldots,r}\{|S^l| : S \subseteq S^l\}$ then the number of variables included will be larger by no more than a constant factor. Depending on how informative the ordering π is, an improvement in the estimation rate may be achievable with such knowledge.

This approach can be straightforwardly applied to the setting where there is only a partial ordering over the variables. One strategy is simply to break any ties randomly and use the ordering that this creates. Alternatively, the grid of subsets can be modified to ensure that variables that are tied are excluded at the same point on the grid.

Since we will subsequently consider procedures that require a choice of tuning parameter λ, it will be useful to expand our notation to use the dual indices $(k, l) \in \{1, \ldots, m\} \times \{1, \ldots, r\}$ for the candidate models. Here the first index specifies the tuning parameter $\lambda^{(k)}$ and the second the subset of variables S^l.

3.2.2 Application to Lasso regression

A popular approach for estimating β in the high-dimensional linear model is the Lasso (Tibshirani, 1996), where $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, defined

$$\hat{\beta} \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{3.3}$$

for some tuning parameter $\lambda > 0$. There exists a wide literature on how λ should be chosen in the vanilla setting, with both theoretical and data-driven choices available (see e.g. Chapter 6 in Bühlmann and van de Geer (2011)). In Section 3.3 we provide a result applicable to the setting where λ is selected by validation on a test set.

In practice, Lasso solutions are nearly always computed in a pathwise fashion (Efron et al., 2004; Friedman et al., 2010), beginning with an initial solution $\hat{\beta} = 0$ at $\lambda^{\text{init}} = \|\boldsymbol{X}^T\boldsymbol{Y}\|_\infty / n$. Solutions are then computed along a path of decreasing values $\lambda^{\text{init}} = \lambda^{(1)} > \lambda^{(2)} > \ldots > \lambda^{(m)}$,

each time using the previous estimate for $\hat{\beta}$ as a warm initialisation. This is often substantially faster than simply computing the solution at $\lambda^{(m)}$, also aiding computational stability. The speed at which solutions can be computed using this approach is a key attraction of the Lasso.

Our proposal is to compute the solution path for each $S^1, \ldots, S^r$ and then select a solution from the set $\{\hat{\beta}^{k,l} : k \in \{1, \ldots, m\}, l \in \{1, \ldots, r\}\}$. While computing this array of solutions will necessarily take longer than a single path, this increase can be substantially reduced by using the following observations.

Suppose that the solution path consisting of $\hat{\beta}^{1,l-1}, \ldots, \hat{\beta}^{m,l-1}$ has already been computed. Letting $k' = \min\{k \in [m] : \text{supp}(\hat{\beta}^{k,l-1}) \not\subseteq S^l\}$, then for $k = 1, \ldots, k'-1$, $\hat{\beta}^{k,l} = \hat{\beta}^{k,l-1}$, so no further computation is required. Finding k' can be done much more cheaply than computing the full path of solutions. To check whether $k' > k$ (for some k), it is sufficient to check simply whether $\hat{\beta}_j^{k,l-1} \neq 0$ for any $j \in S^{l-1} \backslash S^l$. The finer the grid (i.e. the larger r is), the more computation can be skipped and thus the greater the relative gains of this approach are (we show this in Figure 3.1).

Algorithm 3.1 Algorithm for Lasso regression

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\{S^l\}_{l=1}^r$, $\{\lambda^{(k)}\}_{k=1}^m$, $\lambda_0 > 0$

Set $A(k, 0) = \{1, \ldots, p+1\}$ for $k = 1, \ldots, m$ and $\hat{\beta}^{0,l} = 0$ for $l = 1, \ldots, r$
for $l = 1, \ldots, r$ **do**
  for $k = 1, \ldots, m$ **do**
    if $A(k, l-1) \subseteq S^l$ **then**
      Set $\hat{\beta}^{k,l} = \hat{\beta}^{k,l-1}$ and $A(k, l) = A(k, l-1)$
    else if $\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}^{k-1,l}\|_2 / \sqrt{n} \leq \lambda^{(k-1)} / \lambda_0$ **then**
      Set $\hat{\beta}^{k,l} = \arg\min_{\beta : \beta_{S^{l^c}} = 0} \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda^{(k)} \|\beta\|_1$ using $\hat{\beta}^{k-1,l}$ as initial estimate
      Set $A(k, l) = A(k-1, l) \cup \text{supp}\hat{\beta}^{k,l}$
    else
      Set $\hat{\beta}^{k,l} = \hat{\beta}^{k-1,l}$ and $A(k, l) = A(k-1, l)$
    end if
  end for
end for

Output: $\{\hat{\beta}^{k,l} : k \in \{1, \ldots, m\}, l \in \{1, \ldots, r\}\}$

Further time can be saved by observing that the bulk of computation occurs towards the end of the solution path, corresponding to the smallest values of λ. Early stopping can overleap this; `glmnet` (Friedman et al., 2010) opts for a criterion based on the change in residual deviance between consecutive values of λ. The square-root Lasso (Sun and Zhang, 2012) (see also van de Geer (2016)) provides KKT conditions for the Lasso procedure when simultaneously estimating β and the error variance σ, proposing terminating computation

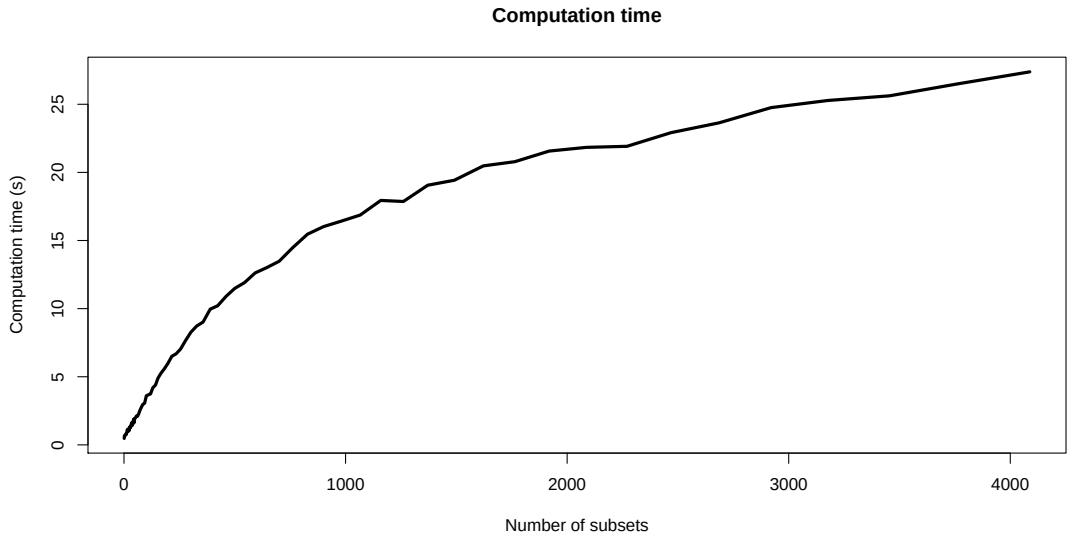


Figure 3.1 Computation time is visibly sublinear in the number of solution paths. Computed on *Riboflavin* dataset, where $n = 71$ and $p = 4088$.

once

$$\|\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}\|_2/\sqrt{n} > \lambda/\lambda_0,$$

for some suitable choice of λ_0. Throughout Section 3.5 we base our choice on the method in Sun and Zhang (2013). These two measures contribute to a substantial reduction in the effort required to compute a full grid of solutions, as shown in Figure 3.1 when applying this method to the *Riboflavin* dataset (Dezeure et al., 2015), which we revisit in Section 3.5.2. Algorithm 3.1 provides detail of how our approach is implemented, in which the sets $A(k, l)$ record which variables have been included in the solution path for the model S^l for $\lambda \geq \lambda^{(k)}$.

3.2.3 Application to ridge regression

Ridge regression (Hoerl and Kennard, 1970) has long been established as a popular method for fitting linear regression models when the data are noisy or large. Within this subsection we define the ridge estimate $\hat{\beta}$ for tuning parameter $\lambda > 0$, in the linear regression setting where $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}. \tag{3.4}$$

The $\lambda\|\beta\|_2^2$ penalty term encourages shrinkage of the ridge solution, so the solution has less variance but experiences some bias towards zero. This tuning parameter plays much the

same role as in the Lasso (3.3). As with the Lasso example in Section 3.2.2, we consider a grid of models corresponding to different values of λ and the subset of $\{1, \ldots, p\}$ that the models are fitted on.

Using the Sherman–Morrison–Woodbury matrix rank-one update formula, we have that ridge solution (absorbing a factor of n into λ for notational convenience) can be written,

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} &= \frac{1}{\lambda}\boldsymbol{X}^T\boldsymbol{Y} - \frac{1}{\lambda}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}\boldsymbol{Y},\end{aligned}$$

meaning that the predictions of this model on new data $\boldsymbol{Z}$ are computed

$$\boldsymbol{Z}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}\boldsymbol{Y}.$$

Supposing that the matrices $\boldsymbol{Z}\boldsymbol{X}^T$ and $(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}$ have already been computed, consider the task of computing predictions for a model over some smaller set of $p_1 < p$ variables, with design matrices $\boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1}$ and $\boldsymbol{Z}_1 \in \mathbb{R}^{n' \times p_1}$ for training and test observations respectively. For simplicity we will assume $n' = n$, and denote $M_0 = (\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}$. Naïve computation of these predictions involves $O(n^2 p)$ computation for the product $\boldsymbol{X}_1\boldsymbol{X}_1^T$, plus $O(n^3)$ for the matrix inversion. For moderate n, this becomes expensive if it needs to be repeated a number of times.

Reordering the columns of $\boldsymbol{X}$ such that $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$, we can obtain (again from the Sherman–Morrison–Woodbury formula):

$$\boldsymbol{Z}_1\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{X}_1^T + \lambda I_n)^{-1}\boldsymbol{Y} = (\boldsymbol{Z}\boldsymbol{X}^T - \boldsymbol{Z}_2\boldsymbol{X}_2^T) \times [M_0 + M_0\boldsymbol{X}_2(I_n - \boldsymbol{X}_2^T M_0 \boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T M_0]\boldsymbol{Y}, \tag{3.5}$$

where the only matrix inversion required is of a $(p - p_1) \times (p - p_1)$ matrix. This computation requires $O(n^2(p - p_1) + n^2 + n(p - p_1)^2 + (p - p_1)^3)$ operations. In the special case where $p - p_1 = 1$, no further matrix inversion is needed, since $1 - \boldsymbol{X}_2^T M_0 \boldsymbol{X}_2$ is a scalar, leading to the overall cost being $O(n^2)$. Similarly, computation of the solution vector $\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{X}_1^T + \lambda I_n)^{-1}\boldsymbol{Y}$ can be shown to be $O(n^2 + np_1)$ in this case.

This forms the basis of our computational strategy for ridge regression models. As in Algorithm 3.1 for Lasso models, we begin by computing predictions for the full model S^1 for all values $\lambda \in \{\lambda^{(1)}, \ldots, \lambda^{(m)}\}$. This sequence can be computed cheaply using only one singular value decomposition $\boldsymbol{X} = UDV^T$, following from the fact

$$M_0 = (\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1} = U\tilde{D}^{\lambda}U^T,$$

where $\tilde{D}^\lambda$ is a diagonal matrix with entries $\tilde{D}^\lambda_{jj} = (D^2_{jj} + \lambda)^{-1}$. This means that each update of λ requires only $O(n^2)$ computation after the SVD has been computed. For each $k \in \{1, \ldots, m\}$, we then can compute the predictions $\boldsymbol{Z}\hat{\beta}^{k,1}, \ldots, \boldsymbol{Z}\hat{\beta}^{k,r}$ by iteratively applying (3.5). In practice, we do this by iterating rank-one updates $p - p_1$ times instead of applying a single rank $p - p_1$ update. Algorithm 3.2 provides described the approach in detail.

Algorithm 3.2 Algorithm for ridge regression

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n'\times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\{S^l\}_{l=1}^r$, $\{\lambda^{(k)}\}_{k=1}^m$
Compute singular value decomposition $\boldsymbol{X} = UDV^T$
Compute $M_1 = \boldsymbol{Z}\boldsymbol{X}^T$
for $k = 1, \ldots, m$ **do**
  Set $M_0 = (\boldsymbol{X}\boldsymbol{X}^T + \lambda^{(k)} I_n)^{-1} = U\tilde{D}^{\lambda^{(k)}}U^T$
  Set $M_2 = M_1$
  Save $\boldsymbol{Z}\hat{\beta}^{k,1} = M_2 M_0 \boldsymbol{Y}$
  for $l = 2, \ldots, r$ **do**
    for $j \in S^{l-1} \backslash S^l$ **do**
      Set $M_2 = M_2 - \boldsymbol{Z}_j \boldsymbol{X}_j^T$
      Set $M_0 = M_0 + M_0 \boldsymbol{X}_j \boldsymbol{X}_j^T M_0 / (1 - \boldsymbol{X}_j^T M_0 \boldsymbol{X}_j)$
    end for
    Save $\boldsymbol{Z}\hat{\beta}^{k,l} = M_2 M_0 \boldsymbol{Y}$
  end for
end for
Output: $\{\boldsymbol{Z}\hat{\beta}^{k,l} \colon k \in \{1, \ldots, m\}, l \in \{1, \ldots, r\}\}$

Since the predictions are cheaper to compute than the solution vector, time and memory can be saved by storing only the predictions of each of the models on test data in the grid. Afterwards, the full solution vector will be computed only for the selected model. Computation of the SVD is $O(n^2 p)$ and the subsequent computation of the predictions over the grid is $O(mn^2 p)$.

To illustrate the speed with which this enables the models to be computed, we again used the *Riboflavin* dataset (see Section 3.5.2). We trained the models on 50 observations and computed predictions for the remaining 21. The finest possible grid of subsets, $r = p$, was used, over a sequence of 100 values of λ were used. Thus, the total number of models considered was 408 800; the full computation for all of these models on a laptop took only 100 seconds. We obtained mean squared prediction error of 0.318 for the model fitted only on the top 169 nodes, compared to 0.399 for the full model.

3.3 Theory

In this section we consider the high-dimensional linear model, $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, where $\beta \in \mathbb{R}^p$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and with errors $(\varepsilon_i)_{i=1}^n$ that are independent and identically distributed with a

σ-sub-Gaussian distribution. We assume that $\boldsymbol{X}$ has identically distributed rows, with zero mean and covariance matrix Σ.

The target of estimation is $\beta \in \mathbb{R}^p$ and we consider the setting where we have been given m candidate estimators $\hat{\beta}^1, \ldots, \hat{\beta}^m$ from which we must select the one that we think is the 'best'. These candidate estimators could, for example, have been computed by Lasso or ridge regressions (see Sections 3.2.2 and 3.2.3 respectively) on separate training data. However, we consider a general setting with no assumption on how these candidates have been generated (other than that they must be independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$). Note that for simplicity, we have reverted to a single index over the candidate estimators.

We define $\hat{\beta}^* \in \{\hat{\beta}^1, \ldots, \hat{\beta}^m\}$ to be the 'best' estimator of the selection based on some unavailable oracular knowledge. In many settings it will be appropriate to take $\hat{\beta}^*$ to be the estimate which minimises the out-of-sample expected prediction error:

$$\hat{\beta}^* = \underset{b \in \{\hat{\beta}^1, \ldots, \hat{\beta}^m\}}{\arg\min} \; (\beta - b)^T \Sigma (\beta - b). \tag{3.6}$$

In order to choose $\hat{\beta}$, we compare each of the $\hat{\beta}^1, \ldots, \hat{\beta}^m$ on a test set of n observations (with design matrix $\boldsymbol{X}$) and select

$$\hat{\beta} = \underset{b \in \{\hat{\beta}^1, \ldots, \hat{\beta}^m\}}{\arg\min} \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X} b\|_2^2. \tag{3.7}$$

Theorem 3.1 shows that by following this strategy, prediction performance on future unseen data will be not much worse than that of $\hat{\beta}^*$. Its proof is provided in Section 3.8.

Theorem 3.1. *Suppose that $\hat{\beta}$ is defined as in* (3.7) *with $\boldsymbol{X} = Z\Sigma^{1/2}$, where Z is a $n \times p$ matrix, with entries Z_{ij} that are independent ν-sub-Gaussian random variables. Then, for quantities (not necessarily constant) $0 < K_1 < n/\log m - 1$ and $0 < K_2$, with probability at least $1 - 2m^{-K_1} - 2m^{-K_2}$,*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta)\|_2 \leq \frac{1 + \Psi}{1 - \Psi} \|\Sigma^{1/2}(\hat{\beta}^* - \beta)\|_2 + \frac{1}{1 - \Psi} 2\sqrt{2}\sigma\sqrt{1 + K_2}\sqrt{\frac{\log m}{n}}, \tag{3.8}$$

where $\Psi = 2\sqrt{2}\nu(1 + K_1)^{1/4}((\log m)/n)^{1/4}$.

Provided $K_1 \log m$ does not grow too fast, these assumptions will be satisfied for sufficiently large n. This result is applicable in the ultrahigh-dimensional setting, e.g. where $\log(p)/n \not\to 0$, as no assumption on p is required. In practice, the candidate solutions $\hat{\beta}^1, \ldots, \hat{\beta}^m$ will typically not be given and will instead need to be estimated, requiring the splitting of samples into training and test sets.

When considering the procedure described in Section 3.2.1, we note that this result holds regardless of the quality of the ordering. For example, if the ordering is unhelpful

or misspecified, this result states that the resulting prediction performance will only be slightly worse than that of the full model. This is validated empirically for a number of Lasso regression models in Section 3.5.1, in which Figure 3.2 shows that performance is not meaningfully impaired relative to the full model.

In addition to above result on prediction error, the following corollary provides a bound on the estimation error of $\hat{\beta}$.

Corollary 3.2. *Assume the same conditions and definitions as in Theorem 3.1, and let σ_l and κ be the minimum eigenvalue and condition number (respectively) of $\Sigma^{1/2}$. Then, with probability at least $1 - 2m^{-K_1} - 2m^{-K_2}$,*

$$\|\hat{\beta} - \beta\|_2 \leq \kappa \frac{1+\Psi}{1-\Psi} \|\hat{\beta}^* - \beta\|_2 + \frac{1}{\sigma_l} \frac{1}{1-\Psi} 2\sqrt{2}\sigma\sqrt{1+K_2}\sqrt{\frac{\log m}{n}}.$$

It will be practical for selection of $\hat{\beta}$ to be done using K-fold cross-validation, as we do throughout the experiments in Section 3.5. There is limited theoretical understanding of the behaviour of cross-validation for model selection in general (see e.g. Bates et al. (2021) for a recent perspective); that which exists is usually specific to the type of estimator being studied (Chetverikov et al., 2021; Feng and Yu, 2019).

3.4 Extensions

We have established, in a regression model such as (3.3) or (3.4), how potential prior information in the form of an ordering (or partial ordering) over the variables can be used to efficiently improve the prediction performance of models. In genomic analysis this can be useful for problems where mutations occur in some proximity of a central location (Aggarwala and Voight, 2016) but the width of this window is not known. Our approach can be generalised to be suitable for a wider variety of structures, such as if both width and location of the window in the previous example are not known.

Suppose we are given $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ on which to fit a number of Lasso or ridge regression models, and wish to compare the predictions of each of the models on a test set of observations with design matrix $\boldsymbol{Z} \in \mathbb{R}^{n' \times p}$.

The fast computation strategies detailed in Sections 3.2.2 and 3.2.3 rely on fitting models on a decreasing sequence of nested subsets of variables. This suggests the possibility of embedding them within certain larger structures that themselves consist of decreasing subsequences. It will be helpful to formulate such structures as a directed graph $\mathcal{G} = (V, E)$ over the set of submodels $\{S^v : v \in V\}$. For a node v, $\mathrm{pa}(v)$ denotes the parents of v and $\mathrm{ch}(v)$ denotes its children.

One such structure is a nested tree, which is a directed tree such that $w \in \mathrm{ch}(v) \implies S^w \subseteq S^v$. A directed tree is a connected graph such that all nodes have exactly one parent

(except for the root node, corresponding to the full set of variables, which has no parent), and with no cycles. In a nested tree, the unique directed path from the root node to any leaf specifies a decreasing sequence of nested subsets to which our approach in Section 3.2.1 is applicable.

Algorithm 3.3 Computational strategy for Lasso regressions in a nested tree structure

Input: $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n' \times p}$, $\boldsymbol{Y} \in \mathbb{R}^n$, $\{\lambda^{(k)}\}_{k=1}^m$, $\mathcal{G} = (V, E)$ a directed tree, $\{S^v\}_{v \in V}$

Set $v = 0$, the root node
Save $\{\boldsymbol{Z}\hat{\beta}^{k,v}\}_{k=1}^m$ and set $C = \{v\}$
while $C \neq V$ **do**
  while $\mathrm{ch}(v) \not\subseteq C$ **do**
    Choose some $w \in \mathrm{ch}(v) \backslash C$ and set $v = w$
    Save $\{\boldsymbol{Z}\hat{\beta}^{k,v}\}_{k=1}^m$ and set $C = C \cup \{v\}$
  end while
  Set $v = \mathrm{pa}(v)$
end while

Output: $\{\boldsymbol{Z}\hat{\beta}^{k,v} \colon k \in \{1, \ldots, m\}, v \in V\}$

Fitted values from all such models can be computed and stored in a memory efficient manner by following the approach outlined at a high level in Algorithm 3.3 for Lasso regressions; the approach for ridge regression models is very similar and is outlined in Section 3.7. The object $C \subseteq V$ used within the algorithm stores indices of the nodes that have been visited before. This means that the number of large objects that require storage (full solution paths in the Lasso regression case, or $\boldsymbol{Z}\boldsymbol{X}^T$ and $(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}$ matrices for ridge regression) depends on the tree structure only as a linear function of its depth.

3.5 Experiments

In this section we explore the properties of our approach in a range of scenarios, using both simulated and real data. In Section 3.5.1 we consider different levels of 'informativeness' in the orderings, and the effect that this has on the prediction error of the final model. Section 3.5.2 explores the effect of grid size, using two real datasets. Sections 3.5.3 and 3.5.4 again use real data, this time exploring how our approach can be used with missing or corrupted data. Lastly, in Section 3.5.5 we run our method on a dataset to predict avocado prices, illustrating the flexibility of our approach by using it within a time series context.

3.5.1 Quality of ordering

In order to see the effect of different variable orderings on the performance of the model, we sample orderings weighted by a vector ρ of probabilities. For example, a neutral (or

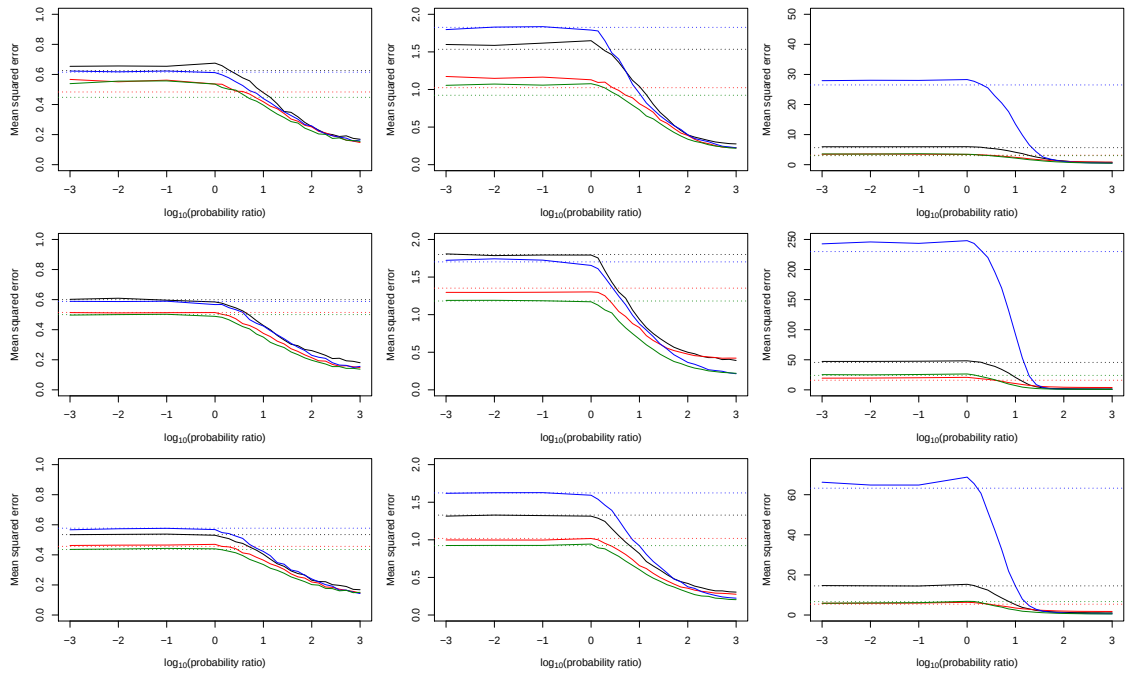


Figure 3.2 Prediction error for varying K: 0 means neutral, larger and negative means a more adversarially bad choice, larger and positive means more informative. Left–Right: 5, 10, 25 signal variables; Top–Bottom: 0.5, 1.5, $U[0,2]$ signal coefficients. Colours for settings are 1. black, 2. red, 3. blue, 4. green. The dotted lines correspond to the errors achieved by the standard approach on the same data.

uninformative) ordering would have this vector as $(1/p, \ldots, 1/p)$ (where p is the number of variables), as all permutations are equally likely.

After a model has been constructed, with a true support S, we specify the j^{th} entry of ρ for $j = 1, \ldots, p$

$$\rho_j = \begin{cases} K/(p + (K-1)|S|) & \text{if } j \in S \\ 1/(p + (K-1)|S|) & \text{otherwise,} \end{cases}$$

where $K > 0$ is the 'probability ratio'. A choice of $K > 1$ means that the ordering is more likely to favour signal variables (meaning the ordering is likely to be useful), whereas $K < 1$ means the ordering will prefer non-signal variables (meaning that the ordering will be actively unhelpful). The vector ρ is used as the weight vector for sampling a permutation which was then used as the ordering.

The design matrices were sampled with $n = 100$ and $p = 1000$, with i.i.d. mean-zero Gaussian rows with covariance matrix Σ. Tests were run with four different choices of Σ:

1. $\Sigma_{jk} = \mathbb{1}_{\{j=k\}}$
2. $\Sigma_{jk} = 0.9^{|j-k|}$

3. $\Sigma_{jk}^{-1} = 0.4^{|j-k|/5}$ $(\approx 0.833^{|j-k|})$

4. $\Sigma_{jk} = 0.5 + 0.5\mathbb{1}_{\{j=k\}}$

As well as this, three different support sizes were used: 5, 10, and 25 variables, with these variables selected uniformly at random. Three regimes for populating the non-zero entries in β were used: two constant (0.5 and 1.5), and one random, where coefficients are drawn $\overset{\text{i.i.d.}}{\sim} U([0,2])$. The tests were repeated 500 times, comparing the prediction performance and computation time against `glmnet`.

It is interesting to note in Figure 3.2 that an adversarially bad ordering does not give rise to any worse performance than a neutral one. There is reason to believe that this should be preferable: in a setting where an ordering is not actively helpful (i.e. it is either neutral or actively bad) then we wish for our procedure to select the full model, S^1. If the ordering is sufficiently bad that the increase in loss for the submodels is larger than the variance of the test error, there is a greater chance that S^1 will be selected.

3.5.2 Riboflavin and prostate data examples

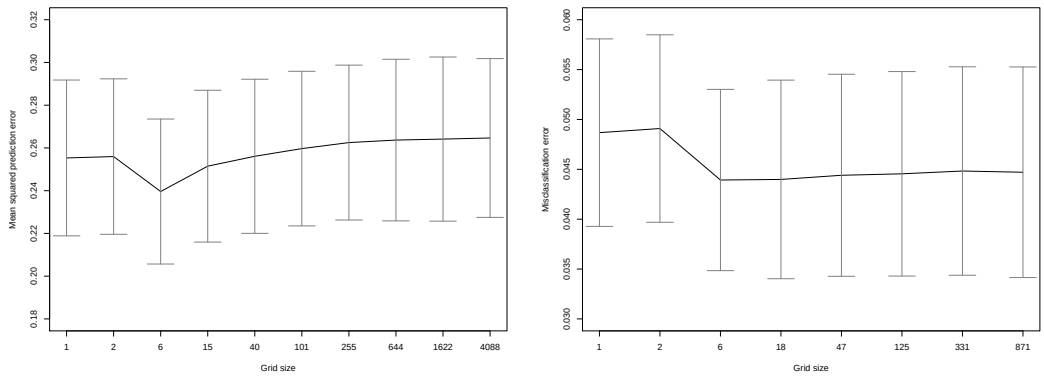


Figure 3.3 Prediction error on riboflavin (left) and prostate (right) datasets for different grid sizes of model. The error bar is ±1 standard deviation.

Tests were performed on two publicly available datasets to explore the performance improvements attained by using this grid-based approach. The two datasets used were the *riboflavin* dataset (available in R package `hdi` (Dezeure et al., 2015)) and the *prostate* dataset (available in R package `spls` (Chun and Keleş, 2010)). Prediction error is estimated by

Dataset	n	p	Response type
Riboflavin	71	4088	Continuous
Prostate	102	6033	Binary

Table 3.1 Summary of dataset information

cross-validation error; within each of the folds cross-validation was also used to select the model. A range of grid sizes were used for each dataset; the ticks on the plots in Figure 3.3 show the sizes used. The ordering used for both of these was the one induced by the scale of the columns in the design matrix.

We see in both of these examples that the prediction error improves after using a grid approach instead of ordinary Lasso models (which is equivalent to a grid size of 1). With the riboflavin data, the error then increases slowly as the grid size increases to the full 4088 subsets, in constrast to the prostate data whose misclassification error remains approximately constant for all grid sizes greater than or equal to 6. Only 8 grid sizes were used for the prostate data due to computational constraints, as for logistic regression models we used a simple loop over `glmnet` rather than the approach described in Section 3.2.2. Each test was repeated 2000 times.

3.5.3 Corrupted data example

As discussed previously, there are many other examples of possible orderings. One such ordering arises when some entries of $\boldsymbol{X}$ are corrupted, which here we take to mean that they are replaced by an i.i.d. standard Gaussian random variable. Here we model this by X_{ij} being corrupted independently with probability ρ_j, which encompasses both homogeneous and heterogeneous corruption settings. In order to obtain an ordering we assume knowledge of the ranking of variables, from most-corrupted to least-corrupted. In practice, if this is not known then an estimate can still be useful, as even if a practitioner has only a very vague notion of which are more likely to be corrupted, we see in Section 3.5.1 that it can still be beneficial to use such an ordering.

We consider the 'muscle-skeletal' dataset from the GTEx project[3], which has 491 rows and 14,713 columns. It is derived from a large-scale ribonucleic acid sequencing experiment from tissue donors, and was provided with a supplementary dataset of known 'confounder' variables, which for this experiment have been regressed out so we use an 'unconfounded' dataset. See Section 4.2.2 for a detailed explanation of the latent confounding model. The response variable was a column randomly selected from the matrix, meaning that for this experiment $n = 491$ and $p = 14713$ after including an intercept term.

We tested performance in four settings, each with a different vector ρ controlling the corruption probabilities of the variables. We construct ρ for each of the four settings as follows (before randomising the order of ρ):

1. $\rho_j = 0.5$ for $j = 1, \ldots, \lfloor 0.2(p-1) \rfloor$ and 0 otherwise
2. $\rho_j = 0.5$ for $j = 1, \ldots, \lfloor 0.5(p-1) \rfloor$ and 0 otherwise

[3]http://gtexportal.org

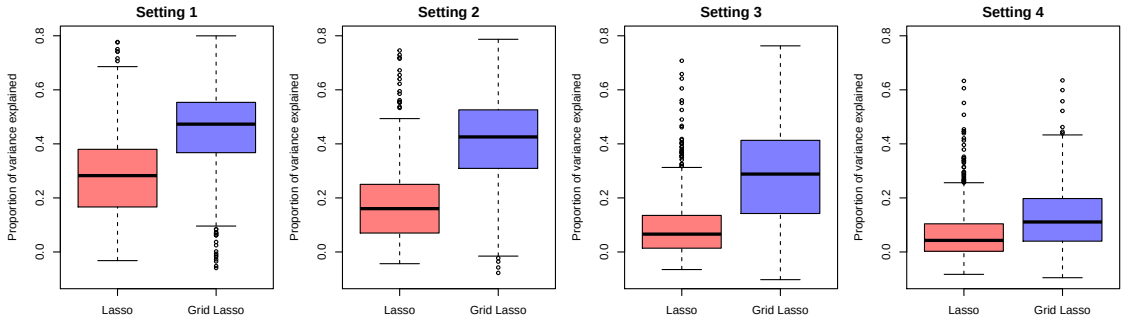


Figure 3.4 Proportion of variance explained, $1 - \text{MSE}/\hat{\text{var}}(\boldsymbol{Y})$, by Lasso and grid-based methods for each of the corruption regimes.

3. $\rho_j = 0.5$ for $j = 1, \ldots, \lfloor 0.8(p-1) \rfloor$ and 0 otherwise

4. $\rho_j = \min\{0.95, (j-1)/(p-1)\}$

A grid size of 25 was used for these experiments and this was repeated 2000 times.

The data were split into five folds; for each of these a model was fitted on the complementary four with entries corrupted according the settings above. These models were themselves tuned using five-fold cross-validation.

3.5.4 Heterogeneous missing data example

Here we consider the missing data setting, where each entry X_{ij} is missing independently with probability ρ_j. In contrast to Section 3.5.3, here we assume it is known exactly which entries are missing and therefore the exact ranking of the variables in terms of their overall missingness is available.

Data are missing homogeneously in the case where $\rho_j \equiv \rho$, i.e. the entries in $\boldsymbol{X}$ are all missing with equal probability. In this case, the probability ρ can be estimated and there are well-studied methods for computing Lasso solutions, such as that discussed in Loh and Wainwright (2012). However, the setting we consider here includes heterogeneous missing data. Within high-dimensional statistics there are methods that accommodate heterogeneous missingness in principal component analysis (Zhu et al., 2019) and in regression problems (Datta and Zou, 2017; Rosenbaum and Tsybakov, 2013).

We begin by observing that the Lasso objective can be written

$$\frac{1}{2n}\boldsymbol{Y}^T\boldsymbol{Y} - \frac{1}{n}\beta^T\boldsymbol{X}^T\boldsymbol{Y} + \frac{1}{2n}\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \lambda\|\beta\|_1, \tag{3.9}$$

which depends on $\boldsymbol{X}$ only through the vector $\boldsymbol{X}^T\boldsymbol{Y}/n$ and matrix $\boldsymbol{X}^T\boldsymbol{X}/n$. In the case where $\boldsymbol{X}$ has some missing entries, the above quantities can be estimated by:

$$\hat{\Gamma}_{jk} = \tilde{\boldsymbol{X}}_j^T\tilde{\boldsymbol{X}}_k/|\{i : X_{ij} \text{ and } X_{ik} \text{ not missing}\}| \tag{3.10}$$

$$\hat{\gamma}_j = \tilde{\boldsymbol{X}}_j^T\boldsymbol{Y}/|\{i : X_{ij} \text{ not missing }\}|, \tag{3.11}$$

where $\tilde{X}_{ij} = X_{ij}$ if X_{ij} not missing, and 0 otherwise. These quantities can be substituted into the update steps in computing solutions to the following surrogate objective

$$\hat{\beta} \in \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{-\beta^T\hat{\gamma}_j + \frac{1}{2}\beta^T\hat{\Gamma}\beta + \lambda\|\beta\|_1\right\}. \tag{3.12}$$

Note that $\hat{\Gamma}$ is not in general positive semidefinite, a requirement for (3.12) to have a finite minimum. In such instances we observe divergent behaviour when computing solutions to this surrogate objective. Work has been done on handling this problem from an optimisation perspective (Loh and Wainwright, 2012). Alternatively the matrix $\hat{\Gamma}$ can be transformed (Datta and Zou, 2017) by projecting it onto $\arg\min_{\Gamma\in\mathbf{S}_+^p}\|\Gamma - \hat{\Gamma}\|_\infty$, which restores the convexity of the problem. However, in practice this is still too computationally intensive for use on larger problems.

We instead transform $\hat{\Gamma}$ to $\hat{\Gamma}_{\text{psd}} = \hat{\Gamma} + \Lambda_{\min}(\hat{\Gamma})I_p$, where $\Lambda_{\min}(\hat{\Gamma})$ is the smallest eigenvalue of $\hat{\Gamma}$. This is equivalent to the addition of a fixed ridge term in the objective:

$$\begin{aligned}\hat{\beta} &\in \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{-\beta^T\hat{\gamma}_j + \frac{1}{2}\beta^T\hat{\Gamma}_{\text{psd}}\beta + \lambda\|\beta\|_1\right\} \\ &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\left\{-\beta^T\hat{\gamma}_j + \frac{1}{2}\beta^T\hat{\Gamma}\beta + \lambda\|\beta\|_1 + \frac{1}{2}\Lambda_{\min}(\hat{\Gamma})\|\beta\|_2^2\right\}.\end{aligned} \tag{3.13}$$

Tests were performed on the same 'muscle-skeletal' dataset as in Section 3.5.3, again randomly selecting a variable each replicate to use as the response. Three missing data regimes were used, each determined by a vector $\rho \in [0,1]^{p-1}$ specifying the independent missingness probability of each variable:

1. $\rho_j = 0.25$ for all j
2. $\rho_j = (j-1)/3(p-1)$, then randomising the order of ρ
3. $\rho_j = 0.3$ for $j = 1, \ldots, \lfloor 0.5(p-1)\rfloor$ and 0 otherwise, then randomising the order of ρ

The data were split into five folds; for each of these a model was fitted on the complementary four with missing entries according to the settings described. These models were themselves tuned using five-fold cross-validation, with scores computed using the estimates $\hat{\Gamma}$ and $\hat{\gamma}$. Experiments were repeated 250 times.

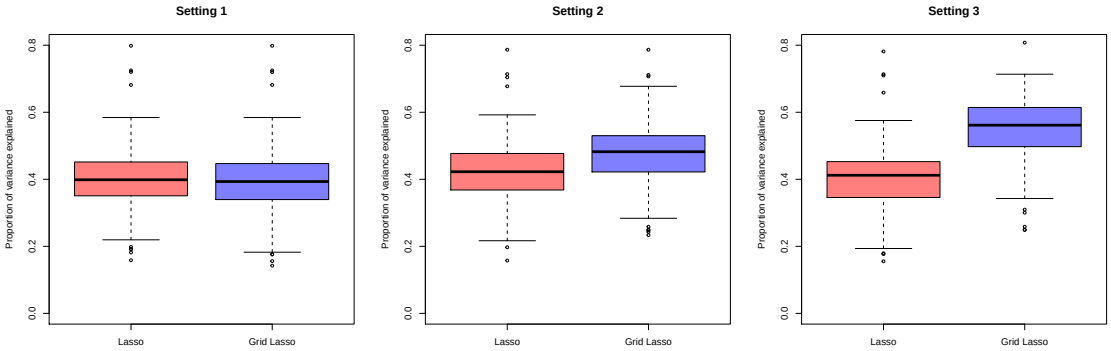


Figure 3.5 Proportion of variance explained, $1 - \text{MSE}/\hat{\text{var}}(\boldsymbol{Y})$, by Lasso and grid-based methods for each of the missing data regimes.

Figure 3.5 displays the proportion of variance explained by the both the vanilla Lasso and of our approach with grid size of 25 in each of the above settings. As expected, there is no improvement in Setting 1 from using a grid approach, as the missingness is homogeneous so the ordering will be uniformly random. In the other two settings, using our approach with the knowledge of which variables are missing more frequently allows us to fit models that provide better predictions.

3.5.5 Avocado data example

Available on *Kaggle* (Kiggins) is a set of historical data on avocado prices and sales volume in multiple US markets. For this experiment we consider only price, and use the 53 markets for which full weekly price data is available for both 'conventional' and 'organic' varieties from the beginning of January 2015 to the end of March 2018.

A design matrix was compiled using all 106 time series, using the previous 52 values for each of one, thus resulting in a 5512 ($= 52 \times 106$)-dimensional model, with 117 observations. For each avocado variety and market, a model was fitted on the first 78 weeks of data and then tested on the remaining 39 weeks to assess performance. Unlike in the other experiments where we tune the model using cross-validation, here we wish to respect the chronological ordering of the observations. We therefore train on the first 39 observations, then validate on the next 39 in order to select the model, falling within the set-up of Theorem 3.1. Once our model is selected we then retrain it on all 78 of the training observations, before testing on the test set consisting of the last 39 observations.

We must now specify the ordering that is to be used when fitting these models. For a univariate time series with no seasonal effects, we would typically order the variables by ranking them from most to least recent (with most recent being the most 'important'). Here

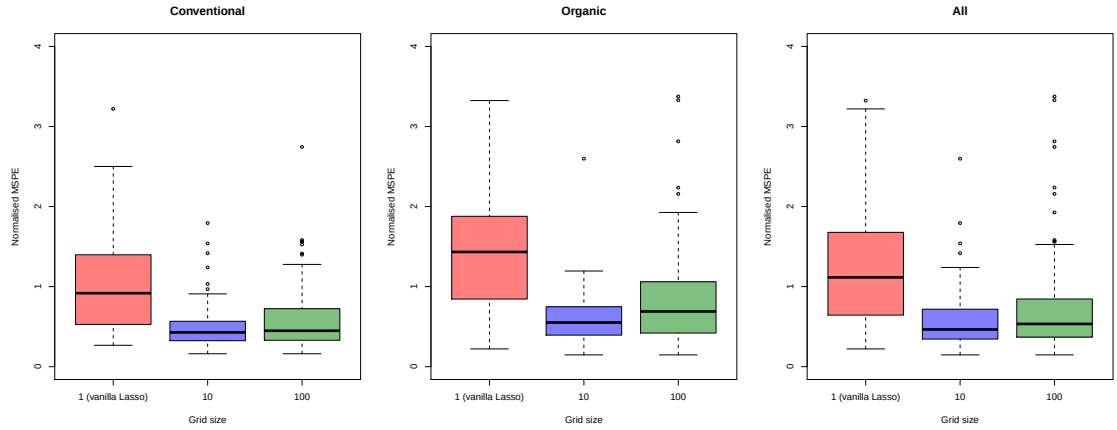


Figure 3.6 Normalised mean-squared prediction error across the models, by grid size. Left and centre plots show the differences between the conventional and organic time series, the right plot includes results for both varieties.

there are effectively 106 time series which are observed weekly, so there may be some seasonal effect.

The ordering used here is constructed as follows, by first splitting the time series into groups of decreasing 'importance':

1. The particular time series that we are modelling
2. The complementary variety to the time series we are modelling (e.g. we are modelling Albany-conventional, here would be Albany-organic)
3. Everything else

Within each of these groups the reading from 52 weeks (one year) previous is first in the ordering, with the rest ordered from newest to oldest.

Each of the response vectors were scaled to have unit variance. Figure 3.6 contains mean-squared prediction error for each time series, fitted with both vanilla Lasso models, as well as our approach with grid sizes 10 and 100. In both cases, using our approach substantially improves the quality of the predictions and that a grid size of 10 gives the largest improvement. This illustrates the flexibility of our approach with respect to the origin of the ordering over the variables, and how it can improve prediction performance in a range of scenarios.

3.6 Discussion

We have introduced a general approach to using potential prior information for improving model fit. We show that one pays at worst only a small theoretical price for using our

approach, but with the potential for significant improvements if the potential prior information is accurate. For two of the most popular and practical methods for fitting high-dimensional linear models, Lasso and ridge regression, we provide computational strategies for fitting the models in a fast and efficient manner. Our experiments demonstrate the performance, efficiency, and flexibility of our approach.

The generality of our approach raises a broad range of questions for extensions and future work. On the theoretical side, there has been recent development in the analysis of cross-validation (Chetverikov et al., 2021) for the vanilla Lasso regression problem. It is possible that related results will be obtainable for cross-validation for selection within more general sets of models, such as we consider in Section 3.3. There is also the potential for further work to extend the result of Theorem 3.1 for generalised linear models.

We provide in Section 3.4 an extension to our framework in order to make use of potential prior information in a nested tree structure. This represents a significant generalisation of our approach; implementation of this strategy to explore its properties empirically will be quite involved. Further still, there is scope for exploring whether our approach can be further extended to other forms of potential prior information relating variables, such as inclusion of interaction terms.

3.7 Supplementary material for Section 3.4

Here we provide details of the small modification to Algorithm 3.3 for use with ridge regression models.

Algorithm 3.4 Computational strategy for ridge regressions in a nested tree structure

Input: $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n'\times p}$, $\boldsymbol{Y} \in \mathbb{R}^{n}$, $\{\lambda^{(k)}\}_{k=1}^{m}$, $\mathcal{G} = (V, E)$ a directed tree, $\{S^v\}_{v\in V}$

for $k = 1, \ldots, m$ **do**
  Set $v = 0$, the root node
  Save $\boldsymbol{Z}\hat{\beta}^{k,v}$ and set $C = \{v\}$
  while $C \neq V$ **do**
    while $\mathrm{ch}(v) \not\subseteq C$ **do**
      Choose some $w \in \mathrm{ch}(v)\backslash C$ and set $v = w$
      Save $\boldsymbol{Z}\hat{\beta}^{k,v}$ and set $C = C \cup \{v\}$
    end while
    Set $v = \mathrm{pa}(v)$
  end while
end for

Output: $\{\boldsymbol{Z}\hat{\beta}^{k,v} : k \in \{1, \ldots, m\}, v \in V\}$

3.8 Proof of Theorem 3.1

We begin with the basic inequality

$$\frac{1}{n}\|\boldsymbol{X}(\hat{\beta}-\beta)\|_2^2 \leq \frac{1}{n}\|\boldsymbol{X}(\hat{\beta}^*-\beta)\|_2^2 + \frac{2}{n}|\varepsilon^T\boldsymbol{X}(\hat{\beta}^*-\hat{\beta})|, \tag{3.14}$$

which follows from the fact that $\|\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}\|_2^2 \leq \|\boldsymbol{Y}-\boldsymbol{X}\hat{\beta}^k\|_2^2$ for all $k \in \{1,\ldots,m\}$. We will first control the second term on the right-hand side of (3.14). Up to and including (3.16), statements are true for any $\boldsymbol{X}$, noting that by definition $\hat{\beta}$ depends on $\boldsymbol{X}$. It follows since the entries of ε are independent σ-sub-Gaussian distributed and independent of $\boldsymbol{X}$ that for probability at least $1-2m^{-K_2}$, we have for each $k \in M$ that

$$\frac{2}{n}|\varepsilon^T\boldsymbol{X}(\hat{\beta}^k-\hat{\beta})| \leq \sigma\sqrt{8(1+K_2)}\sqrt{\frac{\log m}{n}}\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}^k-\hat{\beta})\|_2.$$

In particular, since $\hat{\beta}^* \in \{\hat{\beta}^1,\ldots,\hat{\beta}^m\}$, it follows that

$$\begin{aligned}\frac{1}{n}\|\boldsymbol{X}(\hat{\beta}-\beta)\|_2^2 &\leq \frac{1}{n}\|\boldsymbol{X}(\hat{\beta}^*-\beta)\|_2^2 + \sigma\sqrt{8(1+K_2)}\sqrt{\frac{\log m}{n}}\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}-\hat{\beta}^*)\|_2 \\ &\leq \frac{1}{n}\|\boldsymbol{X}(\hat{\beta}^*-\beta)\|_2^2 + \sigma\sqrt{8(1+K_2)}\sqrt{\frac{\log m}{n}}\left(\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}^*-\beta)\|_2 + \frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}-\beta)\|_2\right).\end{aligned} \tag{3.15}$$

Now, we will use the observation that for $K \geq 0$ and quantities A and B (also non-negative),

$$\begin{aligned} A^2 &\leq B^2 + K(A+B) \\ \implies A^2 - KA &\leq B^2 + KB \\ \implies \left(A-\frac{1}{2}K\right)^2 &\leq \left(B+\frac{1}{2}K\right)^2 \\ \implies A &\leq B + K. \end{aligned}$$

Application of this to (3.15) yields

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}-\beta)\|_2 \leq \frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}^*-\beta)\|_2 + \sigma\sqrt{8(1+K_2)}\sqrt{\frac{\log m}{n}}. \tag{3.16}$$

Now we consider the left-hand side of (3.14). Recall that for a general random variable U that is centred and ω-sub-Gaussian, $U^2 - \mathbb{E}[U^2]$ is sub-exponential with parameters $(32\omega^4, 4\omega^2)$.

Then for $\delta_k \in \mathbb{R}^p$,

$$\begin{aligned}\frac{\delta_k^T(\Sigma - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\delta_k}{\delta_k^T\Sigma\delta_k} &= v_k^T(I_p - \frac{1}{n}Z^TZ)v_k \\ &= 1 - \frac{1}{n}v_k^TZ^TZv_k \\ &= 1 - \frac{1}{n}\sum_{i=1}^n W_i^{(k)^2}\end{aligned}$$

where $v_k = \Sigma^{1/2}\delta_k/\sqrt{\delta_k^T\Sigma\delta_k}$ and thus $\|v_k\|_2 = 1$. We also use $W_i^{(k)} = (Zv_k)_i$, for which we have that each of the n entries of $W^{(k)}$ are independent, centred, ν-sub-Gaussian with $\mathbb{E}W_i^{(k)^2} = 1$.

Thus, $1 - \frac{1}{n}\sum_{i=1}^n W_i^{(k)^2}$ is a centred sub-exponential random variable with parameters $(\frac{1}{n}32\nu^4, \frac{1}{n}4\nu^2)$ by the additivity property of independent sub-exponential random variables, so

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{\delta_k^T(\Sigma - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\delta_k}{\delta_k^T\Sigma\delta_k}\right| \geq t\right) \leq 2\exp\left(-\frac{tn(t\wedge 8\nu^2)}{64\nu^4}\right) \\ \implies \quad &\mathbb{P}\left(\max_{k\in\{1,\ldots,m\}}\left\{\left|\frac{\delta_k^T(\Sigma - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})\delta_k}{\delta_k^T\Sigma\delta_k}\right|\right\} \geq t\right) \leq 2m\exp\left(-\frac{tn(t\wedge 8\nu^2)}{64\nu^4}\right).\end{aligned}$$

We then have by setting $\delta_k = \hat{\beta}^k - \beta$ for $k = 1, \ldots, m$, that with probability at least $1 - 2m^{-K_1}$ for some constant $K_1 > 0$,

$$|(\hat{\beta}^k - \beta)^T(\Sigma - \boldsymbol{X}^T\boldsymbol{X}/n)(\hat{\beta}^k - \beta)| \leq 8\nu^2\sqrt{1+K_1}\sqrt{\frac{\log m}{n}}(\hat{\beta}^k - \beta)^T\Sigma(\hat{\beta}^k - \beta).$$

In the above we have used the assumption that $K_1 + 1 < n/\log(m)$ which implies that $(t \wedge 8\nu^2) = t$ for our choice of t. On this event it follows that for each k,

$$\left|\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}^k - \beta)\|_2 - \|\Sigma^{1/2}(\hat{\beta}^k - \beta)\|_2\right| \leq 2\sqrt{2}\nu(1+K_1)^{1/4}\left(\frac{\log m}{n}\right)^{1/4}\|\Sigma^{1/2}(\hat{\beta}^k - \beta)\|_2,$$

which gives the following inequalities

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta} - \beta)\|_2 \geq \left(1 - 2\sqrt{2}\nu(1+K_1)^{1/4}\left(\frac{\log(m)}{n}\right)^{1/4}\right)\|\Sigma^{1/2}(\hat{\beta} - \beta)\|_2, \tag{3.17}$$

$$\frac{1}{\sqrt{n}}\|\boldsymbol{X}(\hat{\beta}^* - \beta)\|_2 \leq \left(1 + 2\sqrt{2}\nu(1+K_1)^{1/4}\left(\frac{\log(m)}{n}\right)^{1/4}\right)\|\Sigma^{1/2}(\hat{\beta}^* - \beta)\|_2. \tag{3.18}$$

Combining these with (3.16) gives that with probability at least $1 - 2m^{-K_1} - 2m^{-K_2}$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta)\|_2 \leq \frac{1+\Psi}{1-\Psi}\|\Sigma^{1/2}(\hat{\beta}^* - \beta)\|_2 + \frac{1}{1-\Psi}2\sqrt{2}\sigma\sqrt{1+K_2}\sqrt{\frac{\log m}{n}},$$
$$\text{where} \qquad \Psi = 2\sqrt{2}\nu(1+K_1)^{1/4}((\log m)/n)^{1/4}$$

□

Chapter 4

Screening and deconfounding in neighbourhood selection

4.1 Introduction

The familiar regression framework for prediction and estimation considered in the previous chapters is not applicable in settings where there is no natural response variable available. In these scenarios, graphical modelling provides a powerful and intuitive means of understanding the dependence relationships between variables in many settings, such as finance, natural language processing or the modelling of gene regulatory networks (Huynh-Thu et al., 2010; Yin and Li, 2011). Such graphs are useful both for understanding the mechanisms behind various phenomena, and also for enabling predictions to be made based on such a structure. However, in settings where the number of variables is large (potentially much larger than the number of observations), computational considerations become an important factor when selecting a method for graphical estimation.

Suppose that for $i = 1, \ldots, n$, the random vector $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$ has a multivariate normal distribution, with some positive definite covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. These vectors form the rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Even if a variable is only directly related to a small number of others, it can be dependent (in a marginal sense) on a much larger number, so understanding their dependence structure typically requires something other than just finding the marginal dependencies by estimating Σ. This is closely related to the problem of estimating the precision matrix $\Omega = \Sigma^{-1}$ discussed in Section 1.1.2, or the non-zero entries thereof, since for a pair (j, k) of variables,

$$X_j \perp\!\!\!\perp X_k | X_{-jk} \quad \iff \quad \Omega_{jk} = 0. \tag{4.1}$$

This can be shown (Lauritzen, 1996) by deriving the distribution of X_j conditional on observing the remaining entries X_{-j}. The estimates are often presented in the form of

conditional independence graph, whose edges encode the non-zero conditional dependencies between the variables. In the high-dimensional setting of interest where p is larger than n (often much larger), it is not possible to simply compute the empirical covariance estimate and invert it, since this estimate will not be of full rank.

The computational and theoretical aspects of the various existing methods as overviewed in Section 1.1.2 can be highly consequential in practice. For example, constrained ℓ_1-norm minimisation (CLIME) (Cai et al., 2011) enjoys good statistical performance (Loh and Tan, 2018), but is computationally infeasible in high dimensions (Wang and Jiang, 2020) and numerical stability issues have been observed (Zhao et al., 2012). The graphical Lasso (Friedman et al., 2008; Yuan and Lin, 2007) also provides a single optimisation objective for the precision matrix, applying an elementwise ℓ_1 penalty on its entries, and shares similar statistical properties with CLIME. However, it still typically requires $O(p^3)$ operations to solve the problem, despite work to improve the efficiency of the procedure (Witten et al., 2011) for certain sparsity settings. Its computation therefore remains a significant barrier to its use in large-scale settings.

We will focus on the approach of 'neighbourhood selection' (Meinshausen and Bühlmann, 2006; Sun and Zhang, 2013), a procedure based around 'nodewise regressions' as an alternative to estimating the precision matrix via a single large optimisation objective. Recall that these take each column of $\boldsymbol{X}$ in turn as the response variable and perform a Lasso (Tibshirani, 1996) regression (the Dantzig selector (Candes and Tao, 2007) can also be used, as studied in Yuan (2010)) on the remaining columns,

$$\hat{\beta}^j \in \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{4.2}$$

where $\lambda > 0$ is a tuning parameter. The minimiser of the population loss function, β^j, is proportional to $\Omega_{j,-j}$. While the scale of proportionality Ω_{jj} can be recovered by estimating the residual variance in this regression, often the goal is simply to identify which of the off-diagonal entries of Ω are non-zero, in which case this is not necessary. Though generally much faster than the graphical Lasso, neighbourhood selection still involves the computation of p high-dimensional Lasso solutions, which for large p remains a substantial computational task.

An established and effective approach for reducing the computational cost of high-dimensional problems is to reduce the dimension by screening out large numbers of variables beforehand using the approaches mentioned in Section 1.1, such as 'sure independence screening' (SIS) (Fan and Lv, 2008) (also applicable in the generalised linear model setting (Fan et al., 2009)), which thresholds variables based on their sample marginal correlation with the response. They introduce the 'sure screening property', a descriptor of a screening procedure that excludes no true variables with probability approaching 1. Methods such

as this, based on thresholding and screening entries in the empirical correlation matrix, are reliant on the assumption that variables with a large true coefficient have a large marginal correlation with the response, the 'marginal correlation condition', which is not always satisfied. Modifications such as iterating SIS, sequentially fitting models and including further variables marginally correlated with its residuals, and other iterative approaches such as forward regression (Wang, 2009), seek to address this. However the increase in computational cost associated with fitting several times as many models can be large, and these approaches can still suffer in settings where many covariates have moderate to high correlation (Wang and Leng, 2016).

The lack of marginal correlation between conditionally dependent variables remains an issue for neighbourhood selection within the graphical estimation setting; conversely, there can be a large number of highly correlated variables not corresponding to large entries in Ω. We show in Section 4.2.1 that this phenomenon arises naturally in a Gaussian graphical model and thus direct application of correlation screening (Luo et al., 2014) in this setting will generally not possess the sure screening property. We introduce our screening procedure in Section 4.2.3, which is closely related to the first two steps in Xu et al. (2019) which considers a confounding-free setting where the goal is estimation of the full causal structure, and specifically addresses the mechanism by which this can arise within a Gaussian graphical model.

Another significant challenge, both practical and theoretical, to the estimation of Ω is the presence of unobserved confounding variables that affect large numbers of the observed variables in an unknown way. Latent confounders can give rise to large collections of variables apparently highly correlated with one another, as well as spurious conditional dependencies (Novembre et al., 2008) as the confounders are not available for conditioning on. Proper treatment of confounding is important in many applications, such as unobserved environmental factors in scientific experiments (Leek and Storey, 2007) and macroscopic trends in finance (Menchero et al., 2010).

Removal of the top q principal components of $\boldsymbol{X}$ can be an effective option (Fan et al., 2013) if the confounding is well-separated from the signal in the spectrum of $\boldsymbol{X}$, but this will not always be the case. This appraoch also requires correct specification of q, which can be challenging (Barigozzi and Cho, 2020). A number of related approaches for high-dimensional linear regression in the presence of confounders have been developed which involve preconditioning the data (Ćevid et al., 2020; Chernozhukov et al., 2017; Jia and Rohe, 2015) in a way that can each be characterised as spectral transformation of the design matrix $\boldsymbol{X}$ and response $\boldsymbol{Y}$. Removal of the top q principal components is also an example of this approach. After this, a standard procedure such as the Lasso will be applied to these transformed data; see Bühlmann and Ćevid (2020) for details.

To address both problems of computational efficiency and confounding variables, we build on recent advances in covariance estimation to introduce a screening and estimation procedure that is both fast and robust to confounding. Based around the RSVP covariance estimate (Shah et al., 2020), we develop a two-stage screening procedure that specifically addresses the mechanism by which marginally uncorrelated but conditionally dependent variables can arise within a Gaussian graphical model. After obtaining initial screening sets of directly related variables, a 'friends-of-friends' testing step is performed to detect such instances. This same covariance estimate then replaces the empirical covariance in the nodewise regressions (4.2), while also making use of these screening sets, resulting in a fast neighbourhood selection method that is robust to the influence of unobserved confounder variables.

Our approach satisfies the sure screening property under a relaxation of the marginal correlation condition suitable for the graphical estimation setting. Even in a scenario where the sure screening property is not satisfied, provided not too many of the edges are excluded, the performance of the consequent nodewise regressions does not suffer. A range of numerical experiments demonstrate that our proposed approach strikes a favourable balance between accurate estimation of the conditional independence graph and computational efficiency.

4.1.1 Some background on graphical modelling

Graphical models are a convenient and clear way to represent dependence relationships between variables. A graph $\mathcal{G} = (V, E)$ (where V is a set of vertices and $E \subseteq V \times V$ is the set of edges between those vertices) can be either directed or undirected. A popular undirected graphical model is the conditional independence graph, as mentioned in the previous section. This encodes pairwise dependences between variables with the rule that $(j, k), (k, j) \in E$ if and only if $X_j \not\perp\!\!\!\perp X_k | X_{-jk}$.

The subsequent terminology and concepts in this section are not necessary for defining or analysing our procedure. However, they will be useful for the motivation of our approach and these ideas will appear in some of the discussion.

For $j \in V$, the 'adjacency set' of j in $\mathcal{G}$ is defined $\mathrm{adj}_{\mathcal{G}}(j) = \{k \in V : (k, j) \text{ or } (j, k) \in E\}$. For a directed graph, $\mathrm{pa}_{\mathcal{G}}(j) = \{k \in V : k \to j\}$ and $\mathrm{ch}_{\mathcal{G}}(j) = \{k \in V : k \leftarrow j\}$ are the sets of 'parents' and 'children' of j in $\mathcal{G}$ respectively. Similarly, $\mathrm{an}_{\mathcal{G}}(j)$ and $\mathrm{de}_{\mathcal{G}}(j)$ are the 'ancestors' and 'descendants' of j. Note that we use $j \to k$ as shorthand for $(j, k) \in E$.

We consider directed graphical models based on directed acyclic graphs (DAGs), which encode conditional dependence in way more suited to causal interpretation. As is suggested by the name, these are directed graphs in which it is assumed there are no directed cycles, i.e. one cannot have $j \in \mathrm{de}_{\mathcal{G}}(j)$. Such a graph can always be endowed with a 'topological ordering' $\pi \in \mathcal{S}_p$ (the symmetric group of order p) of the nodes, i.e. for $j, k \in V$, if $\pi(k) > \pi(j)$, then there is no directed path from k to j.

We make the assumption that a distribution P, where $X \sim P$, is 'faithful' to some DAG $\mathcal{G}$. This means that for any triplet of disjoint subsets A, B, S,

$$S \text{ d-separates } A, B \quad \iff \quad X_A \perp\!\!\!\perp X_B | X_S,$$

where 'd-separation' (Pearl, 1988) (directed separation) is a graphical criterion for identifying conditional independencies. Sets A and B are d-separated by S if any only if there are no paths (not necessarily directed) in $\mathcal{G}$ between any $j \in A$ and $k \in B$, such that all nodes l on this path *with* $\to l \leftarrow$ either are in S or have a descendant in S, and that all nodes l *without* $\to l \leftarrow$ are not in S. In the case where $A = \{j\}$, $B = \{k\}$ and $S = V \backslash \{j, k\}$, this criterion reduces to the simple rule that j and k are d-separated by $V \backslash \{j, k\}$ if and only if neither of the following hold:

- $k \in \text{adj}_{\mathcal{G}}(j)$
- There exists $l \in V \backslash \{j, k\}$ such that $j \to l \leftarrow k$

For a DAG $\mathcal{G}$, the 'moralised graph' $\mathcal{G}^{\text{m}}$ can be constructed by identifying wherever there exists a triplet $\{j, k, l\}$ such that $j \to l \leftarrow k$ and placing an undirected edge between j and k. Once this is complete, the direction of all of edges are removed to obtain the skeleton graph $\mathcal{G}^{\text{m}}$, which is an undirected graph.

By Proposition 3.25 in Lauritzen (1996) the nodes j, k are d-separated in $\mathcal{G}$ if and only if they are non-adjacent in the moralised graph $\mathcal{G}^{\text{m}}$ of $\mathcal{G}$. Therefore provided that P is faithful to $\mathcal{G}$, by (4.1), for each pair $j \neq k$, $\Omega_{jk} \neq 0$ if and only if $(j, k) \in E(\mathcal{G}^{\text{m}})$. It follows that $\mathcal{G}^{\text{m}}$ is the conditional independence graph for P.

The data-generating process can be modelled by a 'structural equation model' (SEM), which is a set of functions $h_1, \ldots, h_p$ such that for $j = 1, \ldots, p$, $X_j = h_j(\{X_k \colon \pi(k) < \pi(j)\}, \varepsilon_j)$ where ε is a random vector with independent entries and π is a topological ordering over $\mathcal{G}$. A special case of this is a linear Gaussian SEM, where $\varepsilon \sim \mathcal{N}_p(0, I_p)$ and there exists a matrix $A \in \mathbb{R}^{p \times p}$ such that $X = A^T \varepsilon$, where A is upper triangular when row and column indices are ordered according to π.

4.1.2 Organisation of the chapter

In Section 4.2 we begin by discussing the motivation for our approach, in the context of existing screening procedures and the graphical setting. We describe our procedure in detail in Section 4.2.3, including a practical discussion of tuning parameter selection in Section 4.2.4. Section 4.3 provides theoretical support for our procedure, showing that the strong performance of neighbourhood selection is preserved by the use of our screening method. In Section 4.4 we discuss a number of perspectives on settings outside the latent confounding model where we would expect our approach to perform well. We also include a discussion

relating our method to the wider spectral transformation literature and speculate about further research directions. The results from our experiments are contained in Section 4.5, including an exploration of different tuning parameter regimes and a detailed comparison with some alternative approaches.

4.1.3 Notation

For two sequences $(a_n)_{n=1}^{\infty}$, $(b_n)_{n=1}^{\infty}$, we use the notation $a_n \lesssim b_n$ to mean that there exists some constant c such that $a_n \leq cb_n$ for all n. This constant may be universal, or it may depend on other quantities specified as constants in the assumptions (such as in Section 4.2.2) or context (such as A in Theorem 4.2). The statement $a_n \asymp b_n$ means that both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We will also use X to denote a generic random vector, with j^{th} entry the random variable X_j.

4.2 Methodology

In this section we motivate, define, and discuss practical aspects of our procedure. In Section 4.2.1 we expand on some of the discussion in Section 4.1 to provide motivation for our approach. Section 4.2.2 provides details of the two set-ups we consider, without and with confounding, respectively. In Section 4.2.3 we give details of our procedure and a result showing the conditions under which the screening procedure removes no true variables. We also include an upper bound on the number of false positives in each of the screening sets. Section 4.2.4 contains a useful and practical discussion of tuning parameter selection for our approach.

4.2.1 Motivation

Neighbourhood selection (4.2), while quicker than procedures such as the graphical Lasso, can still take a substantial amount of time to carry out. This can be seen, for example, in the experiments in Sections 4.5.3 and 4.5.4. Screening provides a practical and attractive way to dramatically reduce the cost of computing solutions of the nodewise regressions, while remaining sensitive to the true edges.

Algorithm 4.1 Correlation screening / SIS (Fan and Lv, 2008; Luo et al., 2014)

Input: $\hat{\Sigma} \in \mathbb{R}^{p \times p}$, $\tau_1 > 0$
1: **for** $j = 1, \ldots, p$ **do**
2: Set $S_j^{\text{init}} = \{k \neq j : |\hat{\rho}_{jk}| \geq \tau_1\}$
3: **end for**
Output: $\{S_1^{\text{init}}, \ldots, S_p^{\text{init}}\}$

A key issue when applying a screening procedure such as SIS (see Algorithm 4.1) in a graphical setting is that the marginal correlation condition, the existence of some $\omega > 0$ such that $\Omega_{jk} \neq 0 \implies |\rho_{jk}| \geq \omega$, is not generally satisfied. This is also a problem in the vanilla high-dimensional regression setting, where iterative methods have been proposed (Fan and Lv, 2008) to ensure that variables marginally uncorrelated with the response will be eventually included in the screening set. The number of iterations for such procedures to possess the sure screening property can be arbitrarily large; we give an example in Section 4.9 where this is the case.

Within the graphical estimation setting, such iteration can thankfully be avoided, provided that a marginal correlation condition holds on the subset of non-zero off-diagonal entries of Ω that are also edges in the directed graph $\mathcal{G}$ to which we assume the distribution of X is faithful. We noted in Section 4.1.1 that each of the edges (j, k) in the conditional independence graph must belong to one of two categories. We must have either that $j \to k$ or $k \to j$ in $\mathcal{G}$, in which case this is a directed edge, or that there exists a l such that $j \to l \leftarrow k$. The sure screening approach (Luo et al., 2014) can be seen as to require all such edges to correspond to the first case, as then is it reasonable to place a marginal correlation condition on all of the true edges. This is a strong but not unreasonable restriction, as certain graphical structures such as trees will satisfy this.

Figure 4.1 The two possible cases for nodes j and k to be conditionally dependent, given $V \backslash \{j, k\}$. Either there is a directed edge between j and k (left), or there is some node $l \in V \backslash \{j, k\}$ such that $\{j, k, l\}$ form a v-structure (right).

This suggests that in order to identify all of the edges in $\mathcal{G}^{\mathrm{m}}$, one can look at all of the triplets and test for conditional dependence using partial correlation tests (Wille and Bühlmann, 2006). However, this will clearly result in a procedure requiring $O(p^3)$ computational operations which is not suitable for use in a high-dimensional setting. This motivates our screening approach that consists of two stages, the first of which conducts a search for directed neighbourhoods of each node. The second uses these neighbourhoods to only test the conditional independence of 'friends-of-friends', given a common neighbour. This drastically reduces the number of tests that need to be performed, while retaining the ability to identify all of the non-zero entries in Ω.

The possibility of latent confounding can be cleanly addressed in both screening and neighbourhood selection, by making use of the close relationship between covariance estimation and linear regression. This follows from the loss term in the objective for the j^{th} nodewise

regression being equivalently written

$$\frac{1}{2n}\|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\beta\|_2^2 = \frac{1}{2}\left(\hat{\Sigma}_{jj} - 2\beta^T\hat{\Sigma}_{-j,j} + \beta^T\hat{\Sigma}_{-j,-j}\beta\right),$$

for any $\beta \in \mathbb{R}^{p-1}$, thus depending on the observations only through the empirical covariance estimate. This can be replaced with some alternative covariance estimate, such as one robust to latent confounding. This defines a modified neighbourhood selection procedure more suited to this setting, simultaneously compatible with any screening procedure based solely on an estimate of the covariance.

The RSVP estimator (Shah et al., 2020) is able to estimate the covariance in a high-dimensional setting in the presence of latent confounders (see Section 4.2.2 for details), up to an unknown scale factor. This is given by $\hat{\Sigma}^{\text{rsvp}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}^T$; in practice it is computed via a singular value decomposition as described in Section 4.2.2. It is equivalent to computing the empirical covariance with a transformed matrix $\tilde{\boldsymbol{X}} = F\boldsymbol{X}$, where F is the transformation setting all non-zero singular values of $\boldsymbol{X}$ to unit. It is therefore an example of the spectral transformation approach described in Section 4.1, here in the covariance estimation setting. Further discussion of this connection is included in Section 4.4.2.

A related procedure is proposed in Wang and Leng (2016) for screening in the linear regression setting. The 'high-dimensional ordinary least-squares projection' (HOLP) estimator is defined $\hat{\beta} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}$. Their theoretical results are not applicable outside the simplest setting where the covariates are uncorrelated, but the experiments demonstrate promising empirical performance. They do not directly consider the confounding setting but provide intuition that HOLP removes the effect of highly correlated variables (which we discuss in Section 4.4.1), and both highly correlated and factor model examples are included in the experiments. Thanei (2018) shows that the application of HOLP to nodewise regression (for variable j) is equivalent to thresholding entries of $\hat{\Sigma}^{\text{rsvp}}_{-j,j}$:

$$\begin{aligned}
\hat{\Sigma}^{\text{rsvp}}_{-j,j} = \boldsymbol{X}_{-j}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}_j &= \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T + \boldsymbol{X}_j\boldsymbol{X}_j^T)^{-1}\boldsymbol{X}_j \\
&= \boldsymbol{X}_{-j}^T\left((\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1} - \frac{(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1}\boldsymbol{X}_j\boldsymbol{X}_j^T(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1}}{1 + \boldsymbol{X}_j^T(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1}\boldsymbol{X}_j}\right)\boldsymbol{X}_j \\
&= \frac{1}{1 + \boldsymbol{X}_j^T(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1}\boldsymbol{X}_j}\boldsymbol{X}_{-j}^T(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1})\boldsymbol{X}_j, \qquad (4.3)
\end{aligned}$$

where the second line follows from the Sherman–Morrison–Woodbury formula. They argue that HOLP inherits a robustness to latent confounding from its characterisation as the limit of the ridge regression solution as $\lambda \to 0$.

The result (Shah et al., 2020, Theorem 3) that RSVP estimates the underlying covariance matrix Σ elucidates HOLP as a relation of procedures based on thresholding entries of the empirical covariance estimate (Fan and Lv, 2008; Luo et al., 2014). One should therefore

expect it to share their weakness in excluding edges between nodes which are not marginally correlated (before confounding), providing some intuition behind the falsity of the central claim in Wang and Leng (2016) which is known to be incorrect. This motivates the extension of the screening procedure with the two-step strategy in Algorithm 4.2, no longer requiring large marginal correlation between all edges.

4.2.2 Model setting

We will now outline the model setting, introducing the key definitions and concepts that we will use. We assume that the rows of the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ are drawn independently from a multivariate Gaussian distribution,

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma).$$

For simplicity, we will assume that the diagonal entries of Σ are $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$. We slightly abuse notation in assuming that all of the columns of $\boldsymbol{X}$ have been centred such that $\boldsymbol{X}_j = \Pi \boldsymbol{X}_j$, where $\Pi = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the orthogonal projection matrix onto the subspace orthogonal to $\mathbf{1}$, a vector of 1s. This means that an intercept term can be excluded from the nodewise regressions. In the setting without latent confounders (see Section 4.2.2), $\boldsymbol{X}$ is observed directly, and in Section 4.2.2 we introduce the model where we do not observe $\boldsymbol{X}$ directly but instead with latent confounding.

Recalling that the population minimiser of the j^{th} nodewise regression objective is $\beta^j = \Sigma^{-1}_{-j,-j}\Sigma_{-j,j}$, and we define $S_j = \{k : \beta^j_k \neq 0\}$, with $s_j = |S_j|$ being the 'sparsity' of the model and let $s = \max_j s_j$.

Without confounding

In this setting, the design matrix $\boldsymbol{X}$ is observed directly. We will use the empirical covariance estimate $\hat{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ for both screening and the subsequent nodewise regression procedure with the objective:

$$\hat{\beta}^j \in \underset{\beta \in \mathbb{R}^{p-1} :\, \beta_{\hat{S}_j^c} = 0}{\arg\min} \left\{ \frac{1}{2}\beta^T \hat{\Sigma}_{-j,-j}\beta - \beta^T \hat{\Sigma}_{-j,j} + \lambda_j \|\beta\|_1 \right\}. \tag{4.4}$$

In subsequent analysis we will require the following assumption on Σ.

Assumption 1. There exists $c_1 > 0$ such that $\sigma_l = \lambda_{\min}(\Sigma)$ has $\sigma_l > c_1$.

With latent confounding

Now suppose that $\boldsymbol{x}_i$ is not a direct observation of a multivariate normal with covariance Σ, but instead the convolution:

$$\boldsymbol{x}_i = \boldsymbol{w}_i + \Gamma \boldsymbol{h}_i, \tag{4.5}$$

where $\boldsymbol{w}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_p(0, \Sigma)$, $\boldsymbol{h}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}_q(0, I_q)$ and $\Gamma \in \mathbb{R}^{p \times q}$ is a fixed matrix, for $q \geq 1$. Our goal is the same, to estimate $\Omega = \Sigma^{-1}$ from the coefficients in the nodewise regressions, but now we have $\Theta = \operatorname{cov}(\boldsymbol{X}) = \Sigma + \Gamma\Gamma^T$. We will use an estimate of the covariance that is able to target Σ in this setting, RSVP (Shah et al., 2020):

$$\hat{\Sigma}^{\text{rsvp}} = VV^T.$$

Here, V is the matrix of right singular vectors of $\boldsymbol{X}$ obtained via singular value decomposition $\boldsymbol{X} = UDV^T$. We perform screening and nodewise regression with the objective (4.4), using this covariance estimate in place of the empirical covariance. Somewhat more involved assumptions regarding the spectral and geometric properties of Σ and $\Gamma\Gamma^T$ are required for the RSVP estimator to perform well, making use of the following definitions:

$$\begin{aligned} \gamma_u &= \lambda_{\max}(\Gamma\Gamma^T), & \gamma_l &= \lambda_{\min}(\Gamma\Gamma^T), \\ \sigma_u &= \lambda_{\max}(\Sigma), & \sigma_l &= \lambda_{\min}(\Sigma), \\ \rho_1 &= \max_{\|v\|_2=1} \|\Pi_\Gamma \Sigma v\|_2, & \text{and } \rho_2 &= \max_j \|\Pi_\Gamma e_j\|_2, \end{aligned}$$

where e_j is the j^{th} standard basis vector and $\Pi_\Gamma = \Gamma(\Gamma^T\Gamma)^{-1}\Gamma^T$ is the orthogonal projection onto the span of the columns of Γ.

Assumption 2. There exist constants $0 < c_1 < c_2$ such that $c_1 < \sigma_l \leq \max_j \operatorname{var}(X_j) < c_2$, and a constant $c_3 > 1$ such that $\gamma_l > c_3\sigma_u$ and $p > c_3 n$. We also assume that $\log(p) = o(n)$.

Assumption 3. There exist constants $c_4, c_5 > 0$ such that $\sigma_u < c_4 p\sqrt{\log(p)} n^{-3/2} s^{-1/2}$ and $q < c_5\sqrt{n\log(p)/s}$. There also exists a constant $c_6 > 0$ such that $\gamma_u/\gamma_l^2 < c_6\sqrt{\log(p)/sn}$.

Assumption 4. There exists a constant $c_7 > 0$ such that $\max_j {\beta^j}^T \Gamma\Gamma^T \beta^j < c_7$.

Assumption 5. There exist constants $c_8, c_9 > 0$ such that $\rho_1 < c_8$ and $\rho_2 < c_9\sqrt{q/p}$.

In order to understand the performance of $\hat{\Sigma}^{\text{rsvp}}$, it is useful to define another estimator $\hat{\Sigma}^0$ with oracular knowledge of both q and Θ, obtained by setting the q largest eigenvalues of Θ to zero. The conditions in Assumption 2 on the separation of eigenvalues of Σ and $\Gamma\Gamma^T$ are

required for $\hat{\Sigma}^0$ to perform well, and so can be thought of as necessary for the (asymptotic) identifiability of Σ.

The requirement $p > c_3 n$ in Assumption 2 ensures that the characterisation $\hat{\Sigma}^{\text{rsvp}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$ holds, which is used to prove the concentration properties of the estimator. A $p \gg n$ regime is favourable in that, aside from the contribution from the confounding variables, the eigenvalues of the empirical covariance are so noisy as to bear little resemblance to those of Σ, meaning that not much is lost by discarding them. Bounding q ensures that the column space of Γ is not too large, essentially making the contribution from the confounders easier to identify.

In order to prove the results in Section 4.3, there is a so-called 'Gaussian process term' that must be bounded, but this is more involved when it involves a biased covariance estimate. Recovery of the usual estimation error rates for the Lasso requires the entrywise bias of $\hat{\Sigma}^{\text{rsvp}}$ to be controlled at a rate of $\sqrt{\log(p)/sn}$, leading to the conditions given in Assumption 3. The bound for this bias features a $\sigma_u n/p$ term, which will be small in the $p \gg n$ setting with which we are primarily concerned. Within this setting, the conditions in Assumption 3 are less restrictive than they would be for more general problems. Assumption 4 also concerns the control of this Gaussian process term and will hold in realistic settings, such as where the column space of Γ is independently drawn from the uniform distribution over q-dimensional subspaces (Shah et al., 2020). The quantities involved in Assumption 5 can also be adequately controlled within this setting. It is necessary to assume that these are not too large, essentially meaning that the confounding does not overly affect any one of the variables, and is not overly informative of Σ.

It is worthwhile noting that the results in Section 4.3 will still hold under certain relaxations of Assumptions 3–5, but with estimation error rates that no longer match those achievable in the absence of latent confounding.

4.2.3 Our procedure

We will now describe the two-stage screening procedure combined with neighbourhood selection. A single estimate of the covariance or correlation is the only statistic we require from the data. We will use RSVP (Shah et al., 2020) as the covariance estimate on which to base this procedure, but it can be run with the empirical covariance estimate, or any other covariance estimate, which in this section we will denote $\hat{\Sigma}$. The RSVP estimate converges to the true covariance matrix Σ even in the presence of latent confounding, and in the absence thereof does so at the usual $\sqrt{\log(p)/n}$ rate. Further, we give intuition in Section 4.4.1 of additional, related, scenarios where basing the procedure on RSVP can provide advantages.

Algorithm 4.2 Friends-of-friends screening

Input: $\hat{\Sigma} \in \mathbb{R}^{p\times p}$, $\tau_1 > 0$, $\tau_2 > 0$

for $j = 1, \ldots, p$ **do**
  Set $S_j^{\text{init}} = \{k \neq j : |\hat{\rho}_{jk}| \geq \tau_1\}$
end for
for $j = 1, \ldots, p$ **do**
  Set $\hat{S}_j = S_j^{\text{init}}$
  for $k \in S_j^{\text{init}}$ **do**
    for $l \in S_k^{\text{init}} \backslash \hat{S}_j$ **do**
      if $|\hat{\rho}_{jl|k}| \geq \tau_2$ **then**
        Set $\hat{S}_j = \hat{S}_j \cup \{l\}$.
      end if
    end for
  end for
end for

Output: $\{\hat{S}_1, \ldots, \hat{S}_p\}$

Note that, by construction, Algorithm 4.2 is symmetric, i.e. $j \in \hat{S}_k \iff k \in \hat{S}_j$.

Lemma 4.1. *Suppose that for $\omega_1, \omega_2 > 0$ we have an estimate $\hat{\Sigma}$ of the covariance matrix such that there exists some constant $\kappa > 0$ (which need not be known) satisfying*

$$\|\kappa\hat{\Sigma} - \Sigma\|_\infty < \Xi \leq \frac{\omega_2 \lambda_{min}^{(3)}(\Sigma)}{3(8 + \omega_2(1 + \lambda_{min}^{(3)}(\Sigma)))}, \tag{4.6}$$

where $\lambda_{min}^{(d)}(A) = \min_{v:\|v\|_2=1, \|v\|_0 \leq d} v^T A v$ is the minimum eigenvalue of any $d \times d$ principal submatrix of A.

(a) Assume that for any $j \neq k$ such that $\Omega_{jk} \neq 0$, one of the following holds:

(i) $|\rho_{jk}| \geq \omega_1$

(ii) There exists $l \notin \{j, k\}$ such that $|\rho_{jl}| \geq \omega_1$, $|\rho_{kl}| \geq \omega_1$ and $|\rho_{jk|l}| \geq \omega_2$

Then the output of Algorithm 4.2 with selection thresholds (τ_1, τ_2) satisfying $\tau_1 \leq (1 - \Xi)(\omega_1 - \Xi)$ and $\tau_2 \leq 3\omega_2/4$ satisfies

$$\hat{S}_j \supseteq S_j, \quad \text{for } j = 1, \ldots, p. \tag{4.7}$$

(b) Further, define:

$$\mathcal{F}_j^{(1)}(\tau_1, \Xi) = \{k \neq j : \Omega_{jk} = 0, \text{ and } |\rho_{jk}| \geq (1 + \tau_1)(1 - \Xi) - 1\}, \tag{4.8}$$

$$\mathcal{F}_j^{(2)}(\tau_2, \omega_2) = \left\{k \neq j : \Omega_{jk} = 0, \text{ and } \max_{l \in V \backslash \{j,k\}} |\rho_{jk|l}| \geq \tau_2 - \omega_2/4\right\}. \tag{4.9}$$

Then under condition (4.6)*, the number of false variables included in the screening set for node j is:*

$$|\hat{S}_j \backslash S_j| \leq |\mathcal{F}_j^{(1)}(\tau_1, \Xi)| + |\mathcal{F}_j^{(2)}(\tau_2, \omega_2)|. \tag{4.10}$$

Remark 1. The assumptions (i) and (ii), as well as Algorithm 4.2 itself, are motivated by the setting where the distribution P is faithful to some DAG $\mathcal{G}$. In this case, these assumptions is implied by the more interpretable condition in terms of $\mathcal{G}$ where we assume

(i') $(j,k) \in E(\mathcal{G}) \implies |\rho_{jk}| \geq \omega_1$

(ii') $\{(j,l),(k,l)\} \in E(\mathcal{G}) \implies |\rho_{jk|l}| \geq \omega_2$

By the assumption that P is faithful to $\mathcal{G}$, we have that the set of d-separations is exactly the set of conditional independences (see the discussion in Section 4.1.1).

Remark 2. The proof of Lemma 4.1 relies on the fact that from (4.6), the distance between the estimated pairwise correlations $\hat{\rho}_{jk}$, and tripletwise partial correlations $\hat{\rho}_{jk|l}$, and their population counterparts can be uniformly bounded. Thus, we can provide this upper bound on the number of false edges included in the screening sets resulting from Algorithm 4.2 without the need for multiple testing corrections.

We contrast this approach with that used for Theorem 3 in Luo et al. (2014) to control the false positive rate of their procedure. In particular, that requires the maximum correlation between false edges to be of asymptotically strictly smaller order than the minimum correlation between true edges. The results of Lemma 4.1 mean that Algorithm 4.2 can select the correct edges exactly, while only requiring this difference to be of constant order. Our uniform approach to computing the test statistics also means that conservative multiple testing corrections required for related problems, see e.g. Wille and Bühlmann (2006), can be avoided.

We then perform neighbourhood selection (Meinshausen and Bühlmann, 2006) using the screening sets $\{\hat{S}_1, \ldots, \hat{S}_p\}$ we have obtained. For each variable $j = 1, \ldots, p$, we specify penalty parameter $\lambda_j > 0$ and set

$$\hat{\beta}^j \in \underset{\beta : \beta_{\hat{S}_j^c} = 0}{\arg\min} \left\{ -\hat{\Sigma}_{j,-j}\beta + \frac{1}{2}\beta^T \hat{\Sigma}_{-j,-j}\beta + \lambda_j \|\beta\|_1 \right\}. \tag{4.11}$$

These nodewise regressions can be computed very efficiently since $|\hat{S}_j| \ll p$ for each j, with memory consumption dominated by the storage of $\hat{\Sigma} \in \mathbb{R}^{p \times p}$ (which is unavoidable). Section 4.2.4 contains a discussion on selection of λ_j as well as other tuning parameters. We then decide on some threshold, and entries of $\hat{\beta}^j$ of magnitude larger than this threshold are taken to be edges. For simplicity, in Section 4.5 we take this threshold to be zero.

The choice of 'and'/'or' rule for constructing the conditional independence graph (whether $(j,k) \in E \iff \hat{\beta}_j^k \neq 0$ and/or $\hat{\beta}_k^j \neq 0$) remains for the practitioner, as is the case with standard neighbourhood selection. In both practice and theory this choice is not very consequential. Since we are concerned with the post-screening nodewise regressions themselves, rather than the subsequent construction of the conditional independence graph, we consider the results of each of the regressions individually and do not need to consider such a rule.

4.2.4 Parameter selection

Each thresholding step in line 2 and line 8 of Algorithm 4.2 is equivalent to performing a hypothesis test for correlation, or partial correlation, respectively. All of these tests are performed with the same sample size, n, so specifying significance thresholds p_1 and p_2 for these tests is no more than an alternative parameterisation of τ_1 and τ_2. Given n, p_1 and p_2, one can straightforwardly compute the correlation or partial correlation thresholds τ_1 and τ_2 as used in the definition of the algorithm.

In a setting where the available computational budget is approximately known beforehand, it may be preferable to specify $\bar{S}^{\text{init}} = \frac{1}{p}\sum_{j=1}^{p} |S_j^{\text{init}}|$ instead of p_1. This controls the average size of the initial screening set sizes, and in this regime we denote this threshold $\bar{s}_1$. The size of the final sets $|\hat{S}_j|$ cannot be chosen directly, but specifying the size of the initial screening sets will typically provide enough control over computation time of the full procedure.

We note that while these tests are parameterised by their Type I error, Type II error is in fact our primary concern, since typically a false negative in our screening procedure materially worsens our model and a single false positive only slightly worsens computation. Therefore, we argue that multiple testing considerations for the second step of Algorithm 4.2 are not necessary. We expect that often for a pair j, k, with $\Omega_{jk} \neq 0$, there may only be one such $l \in V\backslash\{j,k\}$ with $\rho_{jk|l} \neq 0$, and we wish to ensure that our screening procedure does not exclude this. In a scenario where neighbourhood selection is not performed after the screening step then direct control of false positives may be more important; there has been some investigation of this in e.g. Wille and Bühlmann (2006); Xu et al. (2019).

In Remark 2 it was discussed that for a given pair j, k, the set of test statistics $\{\hat{\rho}_{jk|l} : l \in V\backslash\{j,k\}\}$ typically are highly dependent. This is because they are calculated from the a single covariance estimate $\hat{\Sigma}$, with formula:

$$\hat{\rho}_{jk|l} = \frac{\hat{\rho}_{jk} - \hat{\rho}_{jl}\hat{\rho}_{kl}}{\sqrt{1-\hat{\rho}_{jl}^2}\sqrt{1-\hat{\rho}_{kl}^2}},$$

where e.g. $\hat{\rho}_{jk} = \hat{\Sigma}_{jk}/\sqrt{\hat{\Sigma}_{jj}}\sqrt{\hat{\Sigma}_{kk}}$.

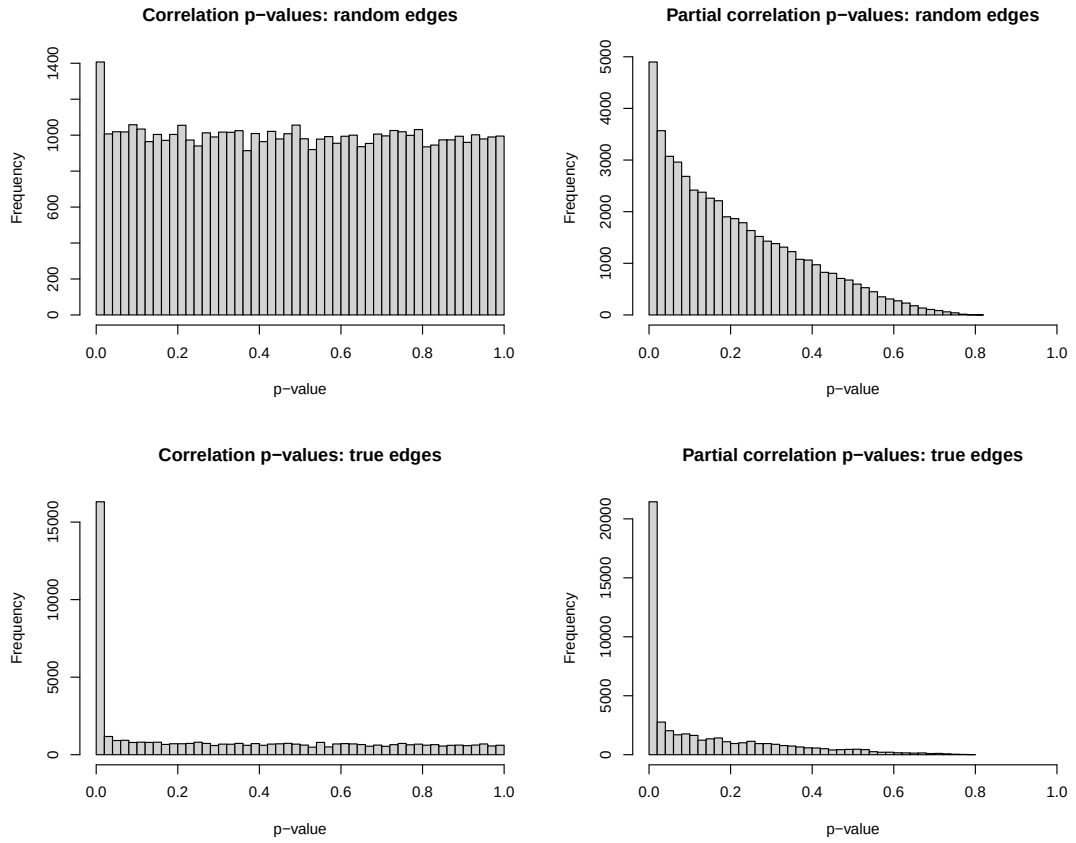


Figure 4.2 Histograms of p-values for experimental setting 5 in Section 4.5.1 with $p = 2000$. Results were similar for the other settings. These are based on a sample of 50 000 true edges (i.e. pairs j, k such that $\Omega_{jk} \neq 0$) and 50 000 edges chosen uniformly at random.

This intuition is validated in practice. Figure 4.2 shows the distribution of p-values for both random and true edges in one of the settings defined in Section 4.5.1, all of which gave similar results. The partial correlation 'p-values' for a pair j, k are taken to be the *minimum* over all $l \in V \backslash \{j, k\}$, which would be very non-conservative in a setting with independent or only slightly dependent p-values. However, we see in the top right panel that the distribution of these 'p-values' for random pairs is only slightly sub-uniform. Indeed, it appears that the familywise error rate (FWER) could be controlled at the nominal level with a corrective factor of approximately 4, rather than the much more conservative factor of 1998 ($= p - 2$) of the Bonferroni correction. In Section 4.5.2 we explore the behaviour of different choices of the ratio p_1/p_2. For simplicity we choose $p_1 = p_2$ and the results in Figure 4.6 suggest that our approach is not highly sensitive to this choice of scaling.

For each of the nodewise regressions (4.11), a choice of tuning parameter λ_j must also be made. Cross-validation (see e.g. Hastie et al. (2009), and the recent paper of Chetverikov et al. (2021) for a theoretical exploration) remains a practical option, but ultimately this

aims to select the optimal model for prediction rather than coefficient estimation, and will include too many non-signal variables (Meinshausen and Bühlmann, 2006). One can also select the parameter with an automatic procedure such as the square-root Lasso (Sun and Zhang, 2012, 2013), which simultaneously estimates the coefficient vector and the noise level, the latter which is used to scale the tuning parameter. In Section 4.3 we will use a universal choice λ for all nodes.

In practice (Friedman et al., 2010), rather than fixing λ_j in advance, solutions are typically computed along a path beginning with large λ_j which then decreases, adding in variables sequentially. This pathwise approach proves much faster and computationally stable than computing a solution at a single point. In our graphical estimation setting the question of explicitly selecting λ_j can be avoided by terminating this sequence of solutions when some criterion is satisfied, e.g. as soon as a certain number of variables have been included, as we do in Section 4.5.3.

4.3 Theory

We will first present results for our procedure assuming that the setting where the screening approach is able to identify all of the true variables. Results will be stated for the RSVP-based method in the setting with potential latent confounding described in Section 4.2.2. The same results also hold (with modified assumptions) in the unconfounded setting of Section 4.2.2 when using empirical covariance based methods. The proofs of both versions of the results are included in Section 4.8. We include a corollary of these results, providing a minimum signal strength condition that ensures exact recovery of the conditional independence graph with high probability.

Theorem 4.2. *Consider the setting in Section 4.2.2 and suppose that Assumptions 2–5 hold and that* $s^2 \log(p)/n \to 0$. *Set* $\lambda = A\sqrt{n \log(p)}/p$, *where* $A > 0$ *and suppose that the conditions of Lemma 4.1(a) hold with*

$$\frac{\sigma_l}{32s} \leq \frac{\omega_2 \lambda^{(3)}_{min}(\Sigma)}{3(8 + \omega_2(1 + \lambda^{(3)}_{min}(\Sigma)))}.$$

With probability at least $1 - c/p$ *for some constant* c, *and for* n, p, A, *sufficiently large,*

$$\max_j \|\hat{\beta}^j - \beta^j\|_1 \lesssim s\sqrt{\frac{\log(p)}{n}},$$
$$\max_j \|\hat{\beta}^j - \beta^j\|_2 \lesssim \sqrt{\frac{s \log(p)}{n}}.$$

Remark 3. The statement of the corresponding result for the unconfounded setting can be obtained by making the following changes to the statement of Theorem 4.2:

(i) Substituting Assumptions 2–5 with Assumption 1,

(ii) Set $\lambda = A\sqrt{\frac{\log(p)}{n}}$ for any choice $A > 0$.

Then the results hold for n, p, sufficiently large.

Theorem 4.2 states that when applying a screening procedure that excludes no signal variables (by virtue of Lemma 4.1), the usual rate of estimation is achieved by the nodewise regression. This is to be expected, as for each of the node the only difference that screening makes is there being a smaller number of nuisance variables.

The following result extends this result to the scenario where the results of Lemma 4.1 do not hold, but nevertheless screening does not exclude too much of the true signal. We specify this in terms of the ℓ_1 norm of the excluded subvector of the population minimiser β^j. Provided this is not too large, the results of Theorem 4.2 can still be achieved.

Theorem 4.3. *Consider the setting in Section 4.2.2 and suppose that Assumptions 2–5 hold and that* $s^2 \log(p)/n \to 0$. *Set* $\lambda = A\sqrt{n \log(p)}/p$, *where* $A > 0$ *and suppose that with probability at least* $1 - \eta$, *for some* $c_{10} > 0$ *and* $j = 1, \ldots, p$ *we have*

$$\max_j \|\beta^j_{\hat{S}_j^c}\|_1 \leq c_{10}\sqrt{\frac{\log(p)}{n}}. \tag{4.12}$$

With probability at least $1 - c/p - \eta$ *for some constant* c, *and for* n, p, A, *sufficiently large,*

$$\max_j \|\hat{\beta}^j - \beta^j\|_1 \lesssim s\sqrt{\frac{\log(p)}{n}},$$
$$\max_j \|\hat{\beta}^j - \beta^j\|_2 \lesssim \sqrt{\frac{s\log(p)}{n}}.$$

Remark 4. The statement of the corresponding result for Setting 4.2.2 can be obtained by making the same changes to the statement of Theorem 4.3 as in Remark 3, additionally replacing (4.12) with $\max_j \|\beta^j_{\hat{S}_j^c}\|_1 \leq \lambda/8$. Then the corresponding result holds for n, p, sufficiently large.

We now include a corollary providing sufficient conditions for the thresholding of the solutions for each nodewise regression to exactly recover the conditional independence graph. It follows directly from the results of one of Theorem 4.2 or Theorem 4.3.

Corollary 4.4. *Assume the set-up of either Theorem 4.2 or Theorem 4.3, and assume there exists a constant C such that for $j = 1, \dots, p$,*

$$\min_{k \in S_j} |\beta_k^j| \geq C\sqrt{\frac{s\log(p)}{n}}.$$

Note that when this is the case, (4.12) *or its empirical counterpart in Remark 4 imply that $\hat{S}_j \supseteq S_j$. Then with probability at least $1 - c/p$ (or $1 - c/p - \eta$, in the case of Theorem 4.3) for some constant $c > 0$, there exists $\tau > 0$ such that for $j = 1, \dots, p$,*

$$\left\{k : |\hat{\beta}_k^j| \geq \tau\sqrt{s\log(p)/n}\right\} = S_j.$$

For simplicity, we have stated the results in this section to be uniform over all j. However, the proofs proceed by showing that the results hold for each of the nodewise regressions, on an intersection of events that occur with high probability. As a consequence of this, the results in fact hold locally; by this we mean that if the assumptions of one of the theorems hold only over a subset of the nodes rather than for all of them, the results are valid for those corresponding nodewise regressions. We include in Section 4.4.1 some discussion of scenarios where this may be the case.

4.4 Possible extensions

4.4.1 Beyond the latent confounding model

We now outline some scenarios outside the scope of the settings introduced in Section 4.2.2. The following are three perspectives on cases which do not fall within the latent confounding formulation (4.5), however there is still reason to expect our RSVP-based approach to perform well.

Correlated predictors. Let us momentarily put aside the graphical construction from which the nodewise regressions arise, and consider simply the regression itself. This brings us closer to the setting in Wang and Leng (2016) whose screening procedure is in this setting closely related to Algorithm 4.1 using RSVP-based covariance estimates.

Consider the case where we have p explanatory variables $j = 1, \dots, p$ and a response $\boldsymbol{Y} = \sum_j \beta_j \boldsymbol{X}_j + \varepsilon$. Here, $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ and $\beta \in \mathbb{R}^p$ is some sparse signal vector with $\|\beta\|_0 = s$. Suppose that the covariance matrix for X has 1 along the diagonal and that all off-diagonal entries are large, positive and similar (but not necessarily equal). We model this scenario for each j,

$$X_j = \alpha_j w + \zeta_j,$$

where w is a standard Gaussian random variable, $\zeta \sim \mathcal{N}(0, Z)$, $Z_{jj} = (1 - \alpha_j^2)$ and the off-diagonal entries $Z_{jk} = \text{cov}(X_j, X_k) - \alpha_j \alpha_k$ are not too large. This describes a high-dimensional regression problem in which all of the covariates are highly correlated. It is known that this scenario can present difficulties for correlation-based screening procedures (Fan and Lv, 2008; Wang and Leng, 2016). Based on the empirical results in Wang and Leng (2016) and the discussion in Section 4.2.1, using RSVP as a basis for screening should provide an advantage.

In this set-up, the full covariance can be computed:

$$\text{cov}([Y, X]) = \left(\begin{array}{c|c} A & B^T \\ \hline B & \text{cov}(X) \end{array}\right), \text{ where} \tag{4.13}$$
$$A = \sigma^2 + \beta^T Z \beta + \left(\alpha^T \beta\right)^2,$$
$$B = Z\beta + \left(\alpha^T \beta\right) \alpha.$$

This suggests the decomposition $\text{cov}([Y, X]) = \Sigma + \Gamma\Gamma^T$, where $\Gamma = (\alpha^T \beta, \alpha) \in \mathbb{R}^{p+1}$ with the remaining covariance

$$\Sigma = \left(\begin{array}{c|c} \sigma^2 + \beta^T Z \beta & \beta^T Z \\ \hline Z\beta & Z \end{array}\right).$$

It then follows that $\lambda_{\min}(\Gamma\Gamma^T) = \|\Gamma\|_2^2 \geq \|\alpha\|_2^2 \gtrsim p$. Provided that conditions on Σ in Section 4.2.2 are satisfied, Theorem 3 of Shah et al. (2020) implies that RSVP applied on $[\boldsymbol{Y}, \boldsymbol{X}]$ will target Σ.

As a consequence, correlation screening based on RSVP will be based on an estimate of $\Sigma_{1,-1} = Z\beta$. Provided Z is close to a diagonal matrix, the non-zero entries of β will correspond to much larger entries in $Z\beta$, making detection of these entries much easier. Contrarily, screening based on the empirical covariance will be based on an estimate of the vector B in (4.13) which is dense since it is equal to $\Sigma_{-1,1}$ plus a multiple of α, whose entries are assumed to be large and approximately equal. This second term is a nuisance and makes detection of large entries of $Z\beta$ more difficult.

Observed confounders. Consider the scenario where $\boldsymbol{X}$ is confounded by a 1-dimensional variable $\boldsymbol{Y}$. While it is known that RSVP estimates Σ when $\boldsymbol{Y}$ is unobserved, we will now see how this behaves when $\boldsymbol{Y}$ is included in the observations. We remark that merely observing the confounding variables is not enough for the empirical covariance (in fact a submatrix thereof) to target Σ; they must be identified as confounders so as to be regressed out. Note that we cannot hope to recover much useful information about the confounding mechanism, since it is dense and therefore the degree of this 'confounder' node can be $O(p)$.

In this case, it follows from the Sherman–Morrison–Woodbury formula that:

$$\begin{aligned}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{Y}\boldsymbol{Y}^T)^{-1}\boldsymbol{X} &= \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X} - \frac{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}}{1+\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}} \\ &= \hat{\Sigma}^{\text{rsvp}} - \frac{\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}}{1+\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}}. \end{aligned} \tag{4.14}$$

The denominator on the RHS of (4.14) is at least 1, from which it follows that:

$$\|\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\boldsymbol{Y}\boldsymbol{Y}^T)^{-1}\boldsymbol{X} - \hat{\Sigma}^{\text{rsvp}}\|_\infty \leq \|\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}\|_\infty^2.$$

Provided $\|\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}\|_\infty$ is small, the observation of $\boldsymbol{Y}$ will have only a small effect on the RSVP estimate. In fact, $\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{Y}$ is exactly the HOLP estimator of Wang and Leng (2016), here in a setting where the predictors $\boldsymbol{X}$ have a dense covariance $Z+\gamma\gamma^T$. Thanei (2018) argues that as this is also $\lim_{\lambda\to 0}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T+\lambda I_n)^{-1}\boldsymbol{Y}$, the limit of the coefficient vector from the ridge regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ as the regularisation parameter $\lambda\to 0$, the entries in HOLP will all be approximately of order $1/p$.

To expand slightly on this intuition, suppose that for each j, we can write (similarly to the previous example of correlated predictors)

$$X_j = \gamma_j Y + \zeta_j,$$

where $\gamma\in\mathbb{R}^p$ is the loading vector for the confounder Y, and $\zeta\sim\mathcal{N}_p(0,Z)$ for some positive definite covariance matrix Z whose off-diagonal entries are small. Then for a small $\lambda>0$, the population ridge objective can be written:

$$\min_{\beta\in\mathbb{R}^p}\mathbb{E}\left[\frac{1}{2}\left(\boldsymbol{Y}-\sum_j \boldsymbol{X}_j\beta_j\right)^2+\lambda\|\beta\|_2^2\right] = \min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\Sigma_{YY}\left(1-\gamma^T\beta\right)^2+\beta^T\left(\frac{1}{2}Z+\lambda I_p\right)\beta\right\}.$$

If we suppose that Z is close to a diagonal matrix, the second term on the right-hand side will have similar shrinkage properties to a ridge penalty (even as we let $\lambda\to 0$). Since γ is a dense loading vector, this optimisation problem will be solved by a vector with many small coefficients, rather than a few large ones.

This intuition is borne out in practice in the following examples. The variables $X_1,\ldots,X_{p-1}$ were set to be highly correlated (but not equally so) with X_p. We then took X_1 as the response variable (within a nodewise regression setting), with s signal variables $X_2,\ldots,X_{s+1}$.

We considered both low-dimensional ($n=200$, $p=100$, $s=3$) and high-dimensional ($n=200$, $p=500$, $s=5$) settings. Three methods were used: nodewise Lasso, ranking of entries in the empirical correlation estimate, and ranking of entries in the RSVP correlation estimate. Each of these return a ranking π of the variables $\{2,\ldots,p\}$, in the case of the

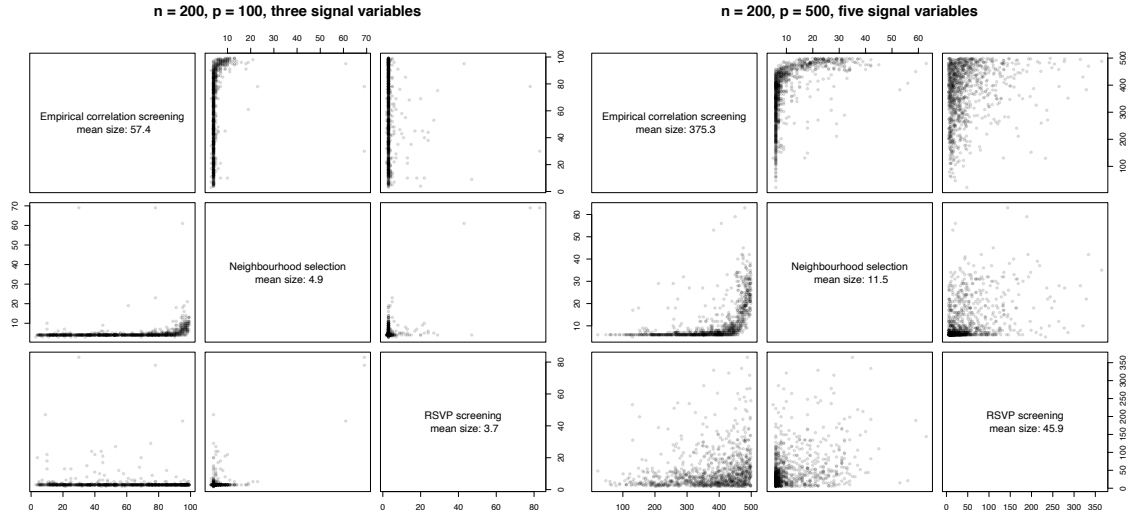


Figure 4.3 Plot of $\pi_{\max}$ for the three variable selection approaches in the two settings.

nodewise Lasso the rank arising from the order by which the variables first take non-zero values on the solution path as λ decreases. We then compare $\pi_{\max} = \max_{k \in \{2,\ldots,s+1\}} \pi(k)$ for each of the methods, with a small value of $\pi_{\max}$ meaning the variable selection approach selects all true variables early on.

Each experiment was repeated 1000 times and results are plotted in Figure 4.3. It is clear that RSVP-based screening far outperforms the approach based on the empirical correlation estimate in both of these examples.

Local sparsity. Sparsity, in the context of neighbourhood selection, is a local concept. By this we mean that it is an assumption imposed nodewise, i.e. for each variable j the population minimiser β^j of its nodewise regression has at most s non-zero entries. In the previous examples, this assumption is valid for all but one of the nodes. For the remaining nodes, where the sparsity assumption holds, one can still typically expect good performance. The nodewise regression approach is not too fragile with respect to the general covariance structure among the covariates, provided e.g. a compatibility condition is satisfied. In this sense, it can be thought of as being robust to non-local violations of the sparsity assumption.

The same robustness to violations of the sparsity assumption is not enjoyed by correlation screening, as we have seen in the previous examples. The application of neighbourhood selection after thresholding the empirical covariance estimate will in general not perform well for *any* of the nodes, since some important variables are likely to be excluded. What is described in the previous example as an observed confounder, may also simply be some variable for which a sparsity assumption is not satisfied. This offers another perspective on the examples where we may expect RSVP-based screening to outperform its empirical counterpart, mirroring the results in Figure 4.3.

4.4.2 Connections with Lava and beyond

Here we discuss the connections between RSVP (Shah et al., 2020) and other spectral transformation approaches, both in the context of nodewise regressions and that of variable screening. In both cases, the RSVP-based approach we use can be characterised as special cases of more general approaches, establishing connections with related methods and inviting potential extensions.

We begin by recalling the Lava procedure (Chernozhukov et al., 2017) which considers the setting where the signal vector can be decomposed into two elements, dense plus sparse. Ćevid et al. (2020) shows that this 'perturbed linear model' is closely related to the latent confounding setting. When applied in a nodewise setting, the Lava objective can be written in the following equivalent forms:

$$(\hat{\delta}, \hat{\theta}) \in \underset{(\delta,\theta)\in\mathbb{R}^{2p-2}}{\arg\min} \left\{ \frac{1}{2n} \|\boldsymbol{X}_j - \boldsymbol{X}_{-j}(\theta+\delta)\|_2^2 + \lambda_1\|\delta\|_1 + \lambda_2\|\theta\|_2^2 \right\}, \tag{4.15}$$

$$\hat{\delta} \in \underset{\delta\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2}\delta^T \boldsymbol{X}_{-j}^T (\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T + 2n\lambda_2 I_n)^{-1} \boldsymbol{X}_{-j}\delta \right.$$
$$\left. - \boldsymbol{X}_j^T (\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T + 2n\lambda_2 I_n)^{-1} \boldsymbol{X}_{-j}\delta + \frac{\lambda_1}{2\lambda_2}\|\delta\|_1 \right\}. \tag{4.16}$$

It can be shown algebraically (Chernozhukov et al., 2017, Theorem 3.1) that these two expressions are equivalent. It is clear that setting $\lambda_1 = \infty$ recovers the ridge regression objective, and setting $\lambda_2 = \infty$ recovers the Lasso objective. In the asymptotic regime $\lambda_2 \to 0$ and $\lambda_1/\lambda_2 \to c > 0$, one recovers the 'Puffer' (Jia and Rohe, 2015) objective in the nodewise setting (4.17), a close relative of the RSVP-based nodewise objective (4.18):

$$\hat{\delta} \in \underset{\delta\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2}\delta^T \boldsymbol{X}_{-j}^T (\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1} \boldsymbol{X}_{-j}\delta - \boldsymbol{X}_j^T (\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T)^{-1} \boldsymbol{X}_{-j}\delta + c\|\delta\|_1 \right\}, \tag{4.17}$$

$$\text{versus } \hat{\beta}^j \in \underset{\beta\in\mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2}\beta^T \boldsymbol{X}_{-j}^T (\boldsymbol{X}\boldsymbol{X}^T)^{-1} \boldsymbol{X}_{-j}\beta - \boldsymbol{X}_j^T (\boldsymbol{X}\boldsymbol{X}^T)^{-1} \boldsymbol{X}_{-j}\beta + c\|\beta\|_1 \right\}. \tag{4.18}$$

In fact any other spectral transformation used for preconditioning a linear regression problem can be used analogously for nodewise regressions. It should be noted however that any theoretical guarantees will generally not be applicable, since their direct application for the j^{th} nodewise regression requires the transformation to be computed based on $\boldsymbol{X}_{-j}$, rather than $\boldsymbol{X}$.

We have discussed in Section 4.2.1 the relationship between HOLP (Wang and Leng, 2016) and variable screening based on thresholding entries of RSVP. Additionally, as the method can be considered from arising from a ridge regression setting (where for consistency we let λ_2 denote the tuning parameter), Wang and Leng (2016) consider the case where λ_2 is

not taken to zero. In a nodewise setting this implies screening based on thresholding the entries of

$$\boldsymbol{X}_{-j}^T(\boldsymbol{X}_{-j}\boldsymbol{X}_{-j}^T + \lambda_2 I_n)^{-1}\boldsymbol{X}_j,$$

which they term 'ridge-HOLP'. By the same argument as (4.3), one can show that this is a scaled version of $\boldsymbol{X}_{-j}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 I_n)^{-1}\boldsymbol{X}_j = \left[\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 I_n)^{-1}\boldsymbol{X}\right]_{-j,j}$. Recall that $\lambda_2 \to 0$ recovers the RSVP estimate, but one can also consider a scaled version as $\lambda_2 \to \infty$:

$$\begin{aligned}\lim_{\lambda_2\to\infty} \frac{\lambda_2}{n}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 I_n)^{-1}\boldsymbol{X} &= \lim_{\lambda_2\to\infty} \frac{1}{n}\boldsymbol{X}^T\left(\frac{1}{\lambda_2}\boldsymbol{X}\boldsymbol{X}^T + I_n\right)^{-1}\boldsymbol{X} \\ &= \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}.\end{aligned}$$

This is the empirical covariance, and can be used within Algorithm 4.2 to provide an effective screening procedure in the setting without latent confounding. Since the methods we outline for the two settings in Section 4.2.2 can be considered to correspond to the cases $\lambda_2 = 0$ and $\lambda_2 = \infty$, we consider in Figure 4.4 the procedure corresponding to some choice $\lambda_2 \in (0, \infty)$. This suggests a nodewise procedure similar to Lava regressions (in the way that (4.17) is similar to (4.18)), with screening based on Algorithm 4.2 substituting $\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 I_n)^{-1}\boldsymbol{X}$ in place of the covariance estimate. A critical obstacle to using such a method is that to our knowledge, it is not known under what conditions $\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda_2 I_n)^{-1}\boldsymbol{X}$ estimates the covariance, or whether it even does at all.

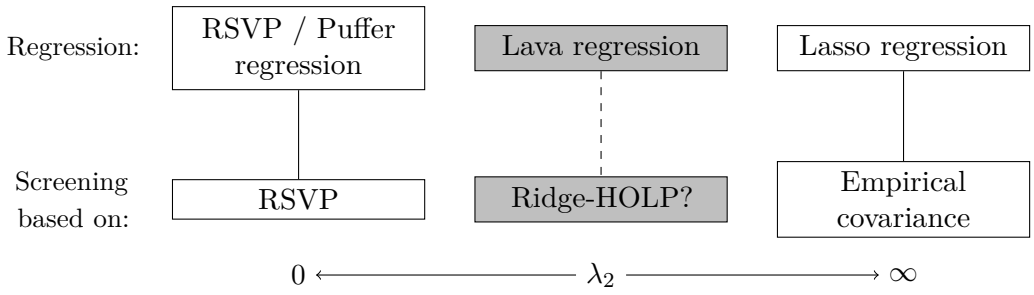


Figure 4.4 Schematic diagram showing the joint screening and estimation procedures specified by different values of λ_2.

This idea can be generalised for a given spectral transformation F of $\boldsymbol{X}$, $\tilde{\boldsymbol{X}} = F\boldsymbol{X}$, such as the 'Trim transform' of Ćevid et al. (2020) which places a ceiling on the singular values of the transformed matrix. There are only certain choices of F for which it is known that $\frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ consistently estimates Σ, such as $F = I$ which gives the empirical covariance estimate and $F = (\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}$ yielding RSVP. An extension of the result of Theorem 3 in Shah et al. (2020) would make the somewhat unifying view of spectral transformations provided in Ćevid et al. (2020) also applicable in covariance estimation setting. This generalised approach would

be immediately compatible with the screened neighbourhood selection procedure introduced in Section 4.2.3.

4.5 Experiments

Here we explore the practical behaviour of our procedure across a range of settings. Sections 4.5.2 and 4.5.3 were based on simulated datasets over a range of settings; for detail of the experimental settings see Section 4.5.1.

In Section 4.5.2 we explore the screening step prior to the neighbourhood selection stage of the process. Continuing from the discussion in Section 4.2.4, we investigate the effect of the choice of the two p-value thresholds, p_1 and p_2. We plot the true positive rate as a function of screening set size, averaged over all p nodes, in order to gain a like-for-like comparison of the different settings. Section 4.5.3 presents the results of a the full procedure described in Section 4.2.3 against a selection of other approaches. Tests were performed across a range of settings, both with and without latent confounding. Estimation accuracy is presented against average computation time, making clear the trade-offs involved with the choice of modelling procedure and tuning parameters. In Section 4.5.4 we see the effect of our variable screening approach on the solution paths on some real data examples, obtained from the GTEx consortium. We consider the similarity of solution paths between those in nodewise regressions with and without having applied our screening method.

4.5.1 Experimental settings

We consider eight scenarios as described below. The covariance Σ is generated by first constructing the precision matrix Ω as specified in each of the settings, then inverting it. This ensures that Ω satisfies the sparsity assumption (see Section 4.2.2). For settings 5–8, the method for including v-structures into an existing precision matrix Ω is described in detail in Section 4.7. Non-zero off-diagonal entries in Ω are set to either -0.3 or 0.7 independently and with equal probability, while ensuring that Ω is symmetric. This is similar to the construction as in Luo et al. (2014)

1. Banded: For $i = 1, \ldots, p$, set Ω_{ij} and Ω_{ji} to be non-zero if $|j - i| \leq 3$ where 3 is the 'bandwidth'.

2. Block model: partition the variables into groups of size 10, if i, j are in the same group let Ω_{ij} and Ω_{ji} be non-zero.

3. Stochastic block model: partition the variables into groups of size 15, if i, j are in the same group let Ω_{ij} and Ω_{ji} be non-zero independently with probability $1/4$, and otherwise let them be non-zero with probability $2/p$.

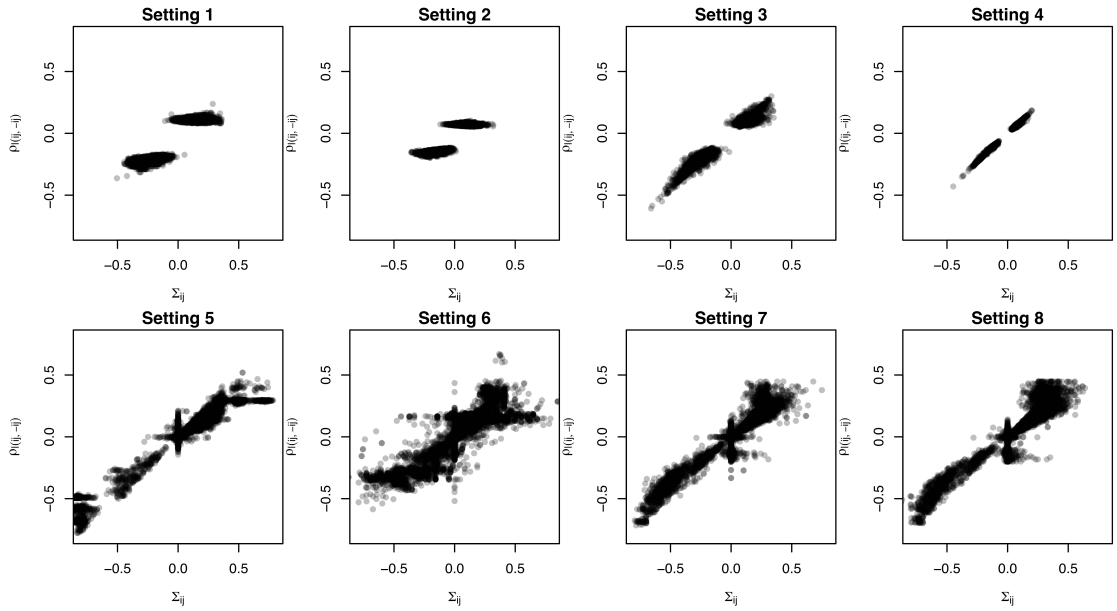


Figure 4.5 Scatter plots of the magnitude of entries of Σ and Ω for edges in the conditional independence graph. For settings 5–8, the addition of v-structures creates in turn other (much weaker) v-structures, corresponding to the points with approximately zero partial correlation.

4. Erdős-Rényi: Let Ω_{ij} and Ω_{ji} be non-zero independently with probability $10/p$.

5. As Setting 1, except with bandwidth 1 and v-structures added between eligible nodes with probability $1/p$.

6. As Setting 2, except with group size 5 and v-structures added between eligible nodes with probability $1/p$.

7. As Setting 3, except with within-group edge probability of 0.1 and otherwise edge probability of $1/p$, and v-structures added between eligible nodes with probability $1/p$.

8. As Setting 4, except with edge probability $2/p$ and v-structures added between eligible nodes with probability $1/p$.

For each of these settings $n = 500$ observations are generated independently, with $p \in \{1000, 2000\}$. Figure 4.5 shows the correlation, and partial correlation, of all pairs of nodes that are adjacent in the conditional independence graph.

4.5.2 Screening set coverage

To explore the coverage of the screening procedure, for the 8 settings described in Section 4.5.1 with $p = 2000$, we ran Algorithm 4.2 with a range of parameters (see Section 4.2.4). For

Algorithm 4.2 we considered five different scaling relationships between p_1 and p_2, with $p_1/p_2 \in \{2^{-2}, 2^{-1}, 2^0, 2^1, 2^2\}$. We also include Algorithm 4.1 which is equivalent to setting $p_2 = 0$.

For each of the final screening sets $\hat{S}_j$, we calculated the true positive rate (the proportion of true edges which are contained in $\hat{S}_j$). These sets are then allocated bins of length 20 according to $|\hat{S}_j|$ and averaged true positive rate is plotted against (binned) screening set size.

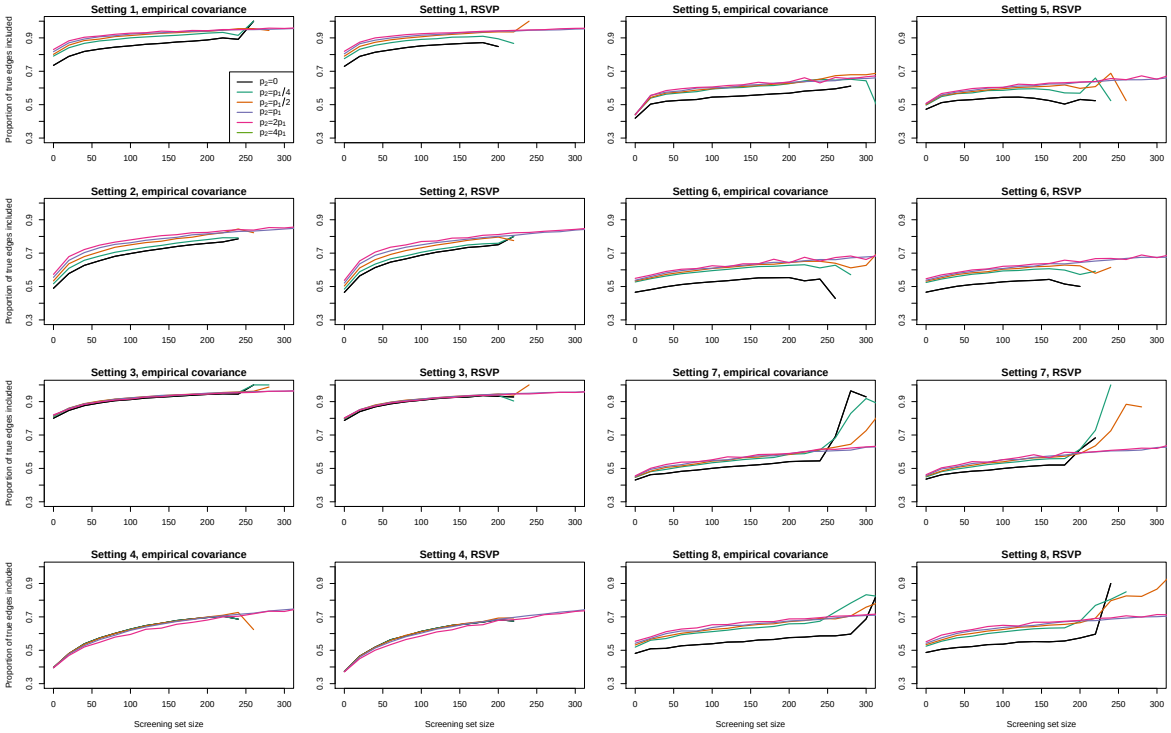


Figure 4.6 Performance of different screening settings as a function of screening set size, averaged over all 2000 nodes.

Figure 4.6 displays the results for screening based on the empirical covariance estimate and also the subsampled RSVP estimate[1]. Some caution is advised in interpreting the results for large screening set sizes (larger than 200 or so) because there were relatively few examples that returned such a set, so results for these sizes will have more variance than elsewhere. We also note that all of these experiments are in a setting with sparse Ω and no latent confounding, and that here the performance of empirical covariance-based and RSVP-based screening is very similar.

Across this range of settings we see that the second stage of the screening process appears to provide better results, with the possible exceptions of settings 3 and 4 where performance

[1] With 100 replicates, each using a subsample of 100 observations. These are the default arguments provided at `https://github.com/benjaminfrot/RSVP`.

is very similar. Remarkably it is not only settings 5–8, with their specific construction of v-structures, where the black line (corresponding to Algorithm 4.1) is visibly lower than the rest. The plots for settings 1 and 2 in Figure 4.5 suggest this may be the case, as visible are points with non-zero partial correlation but approximately zero marginal correlation. Differences between the different values of p_1/p_2 are relatively small, so for simplicity we will make the choice $p_1 = p_2$.

4.5.3 Simulations

Here we explore the properties of neighbourhood selection with our screening procedure as outlined in Section 4.2.3. We measure performance by using the area under the precision-recall curve, which plots Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$ against Recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. Modification of this metric is required since this curve is only completed once Recall $= 1$, but any high-dimensional nodewise Lasso regression can only select at most n variables (Zou and Hastie, 2005). We address this by only plotting the curves for the first 125 variables to appear in the solution path, then computing the area underneath that segment.

We consider settings 1–8 in Section 4.5.1 for the two different model sizes, $p = 1000$ and $p = 2000$. For each setting we also ran experiments with latent confounding present; this followed construction in Shah et al. (2020) with parameter choices $\nu = 1$ and $q = 20$ (matching the choices in constructions (a)-(c)). In particular, the matrix $\Gamma \in \mathbb{R}^{p\times q}$ was populated with i.i.d. standard Gaussian entries, with the k^{th} column then scaled by e^{-k} for $k = 1, \ldots, 20$. Each experiment was repeated 100 times.

In these experiments we include neighbourhood selection (Meinshausen and Bühlmann, 2006) (4.2) without any screening performed. Post-screening neighbourhood selection methods using either Algorithms 4.1 or 4.2 were also included, with $p_1 \in \{0.025, 0.05\}$ (setting $p_1 = p_2$ for Algorithm 4.2 as discussed in Section 4.5.2). As a point of comparison, we included the graphical Lasso (Yuan and Lin, 2007); for computational reasons we are only able to include this for examples where $p = 1000$. Each of these methods only require from the data an estimate of the covariance, so were run using both the empirical covariance estimate and the subsampled version of the RSVP estimate (using 100 subsamples, each containing 100 observations).

We note that as we are only interested in the first region of the solution paths for each of the nodewise regressions, we terminated computation after the path included 250 variables to avoid excess computation. Unfortunately, it was not possible to do this when computing the graphical Lasso solutions, so its computation time will be an overestimate. It is however well-known (Witten et al., 2011) that it is typically substantially slower than neighbourhood selection, so this is very unlikely to have meaningfully affected the results.

Computation time is a key motivation for choosing to screen variables in the first place, so we report in Figure 4.7 the average precision against average computation time for each

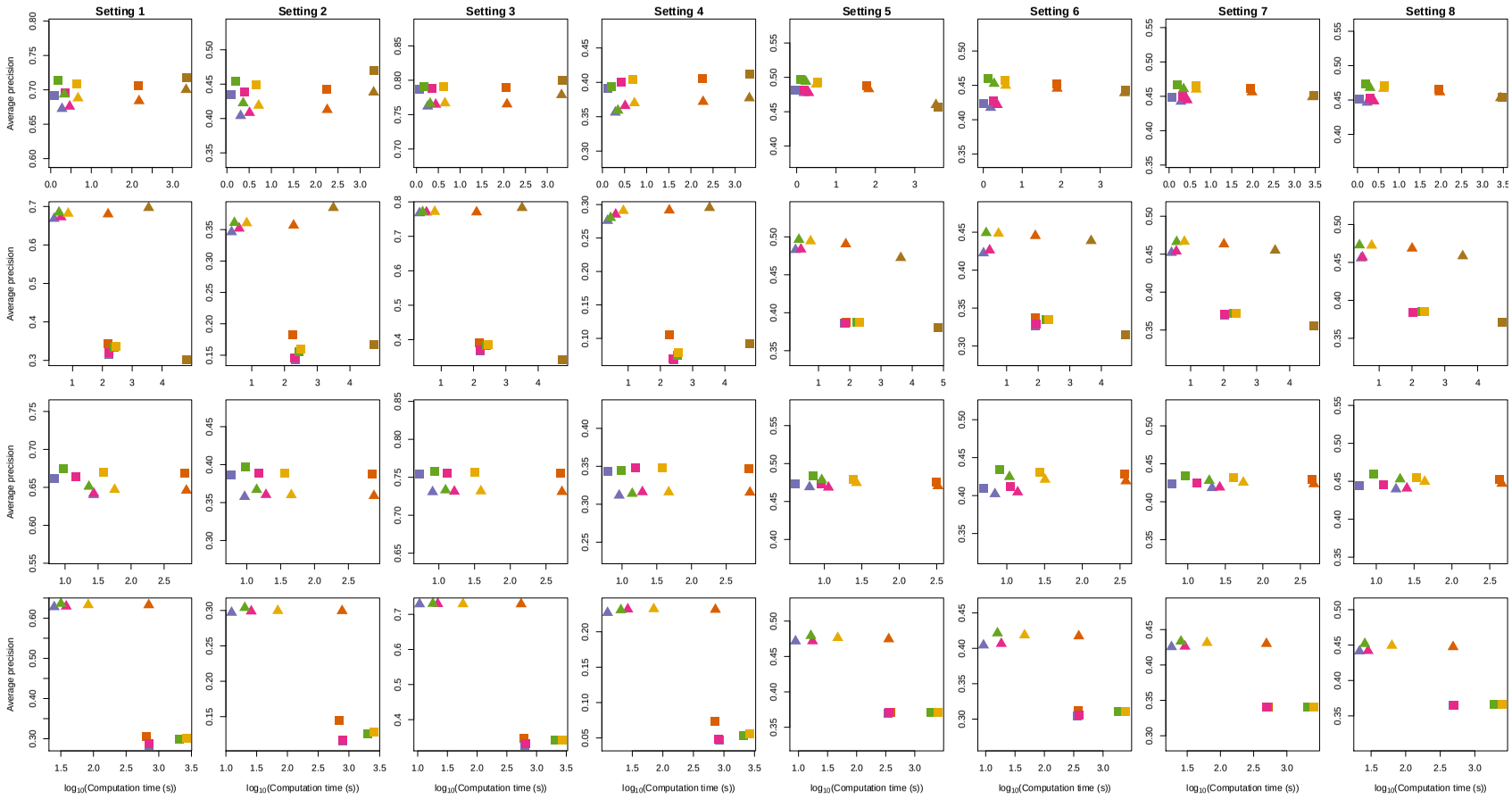


Figure 4.7 Scatter plot of average precision against computation time for the various methods: Full neighbourhood selection (NS) (orange), Post-screening NS with $(p_1, p_2) = (0.025, 0)$ (lilac), Post-screening NS with $(p_1, p_2) = (0.05, 0)$ (magenta), Post-screening NS with $(p_1, p_2) = (0.025, 0.025)$ (green), Post-screening NS with $(p_1, p_2) = (0.05, 0.05)$ (mustard), Graphical lasso (brown). Rows 1–2 have $p = 1000$, rows 3–4 have $p = 2000$. Squares are methods using the empirical correlation estimate, triangles the subsampled RSVP estimate. Rows 2 and 4 are with latent confounding.

of the methods. This shows that screening provides significant speed improvements on full neighbourhood selection, which is in turn faster still than the graphical Lasso. In the absence of latent confounding, the RSVP-based approaches perform almost as well as those which make use of the the empirical covariance estimate, but when confounding is present they have a clear advantage.

Additionally, latent confounding can have the effect of causing a lot of variables to appear highly correlated, meaning that empirical covariance based screening will either fail to include many of the true variables, or include a large number of nuisance variables. Figure 4.7 shows that empirical covariance-based methods require significantly more computation time when confounding is present. The difference in computation time between RSVP and the empirical covariance estimates themselves is negligible, as is the time taken to perform the second stage of screening in Algorithm 4.2 in comparison to the time taken for neighbourhood selection. Across all of the settings, our procedure using RSVP exhibits strong performance, taking a small proportion of the time required for full neighbourhood selection.

4.5.4 Real data examples

The GTEx project[2] (Aguet et al., 2017) provides large-scale data from an extensive ribonucleic acid sequencing experiment. The goal of analysing such data is to identify which genes regulate each other, meaning they are biologically related. It provides gene expression data for samples of 53 different tissues from 714 donors. Datasets are provided fully processed, filtered, and normalised, and for each one supplemental data are provided from known confounding variables. These can be regressed out to provide a 'unconfounded' dataset, though there remains the possibility of further unknown confounders.

From this collection, we randomly selected four datasets with over 250 observations. These had n ranging between 335 and 414, and p between 14,337 and 16,195. Experiments were run on both plain and 'unconfounded' versions of these datasets.

In order to explore the behaviour of our screening approach on real data, on each dataset we performed full neighbourhood selection, and then four variants of our screening approach (all using RSVP). As discussed in Section 4.2.4, instead of specifying the threshold τ_1 via a p-value p_1, for computational reasons we instead specify an average size $\bar{s}_1 \in \{250, 500\}$ for $\bar{S}^{\text{init}}$. For the second screening stage, we used $p_2 \in \{0.025, 0.05\}$, similarly to as in Section 4.5.3. All of these were run using the basic RSVP estimator, without any subsampling.

Each nodewise regression was run until the point where 250 variables were included in the solution path. As the number of variables included in the nodewise regressions increases along the solution path of each nodewise regression, we plot the average (over all p nodes) Jaccard index of the variables selected by the procedures with and without screening at each set size. Table 4.1 shows that substantial amounts of computation time are saved by using

[2] `http://gtexportal.org`

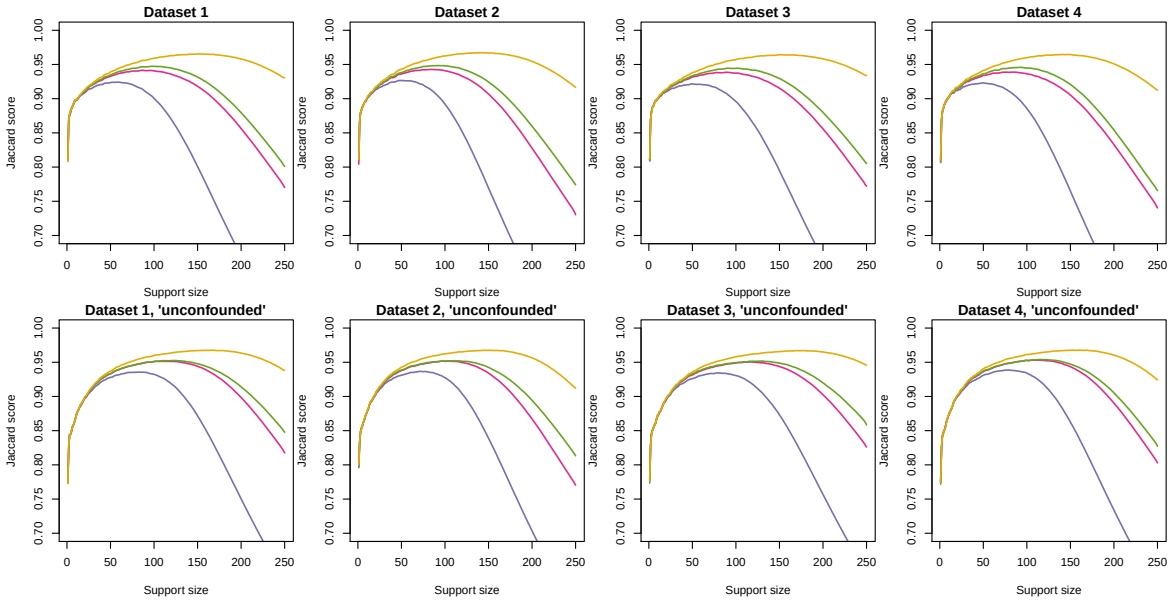


Figure 4.8 Jaccard index of support sets of post-screening neighbourhood selection against full neighbourhood selection. Screening settings: $(\bar{s}_1, p_2) = (250, 0)$ (lilac), $(\bar{s}_1, p_2) = (500, 0)$ (magenta), $(\bar{s}_1, p_2) = (250, 0.025)$ (green), $(\bar{s}_1, p_2) = (500, 0.05)$ (mustard).

screening, while providing similar results in final model (see Figure 4.8) compared to full neighbourhood selection.

Screening settings	Dataset:	1	1-UC	2	2-UC	3	3-UC	4	4-UC
No screening		5869	6299	7572	7103	8214	8061	5168	5301
$(\bar{s}_1, p_2) = (250, 0)$		671	737	967	789	769	675	729	600
$(\bar{s}_1, p_2) = (500, 0)$		693	785	985	900	825	835	725	687
$(\bar{s}_1, p_2) = (250, 0.025)$		825	989	1160	1137	1024	1077	825	827
$(\bar{s}_1, p_2) = (500, 0.05)$		1507	1743	2020	2032	1932	2007	1443	1469

Table 4.1 Computation time (s) of various screening approaches. -UC denotes 'unconfounded' version of the dataset.

4.6 Discussion

In this work we have introduced a new screening approach for neighbourhood selection that reduces computation without overly compromising coverage. The construction of our procedure addresses a common weakness of alternative procedures requiring a marginal correlation condition, within the specific context of conditional independence graph estimation. Our new application of recent developments in covariance estimation with latent confounding enables screening within neighbourhood selection in a wider range of practical settings than is currently available. We have discussed in Section 4.4 how our approach has robustness beyond just the latent confounding framework. Our screening procedure is itself fast, enabling

substantial improvements on the computation time of neighbourhood selection. Through our theoretical results and numerical experiments, we have shown that the cost of doing so in terms of estimation performance is small (if not negligible) both in theory and in practice.

Our assumptions required in Lemma 4.1 for our two-step approach to preserve all of the true edges are usefully weaker than a minimum marginal correlation condition. There is scope for the future development of practical criteria on a structural equation model for these assumptions to be satisfied. This is a challenging question and may entail further work in the direction of notions of faithfulness (see e.g. Meek (1995); Ramsey et al. (2006); Sadeghi (2017)).

In Section 4.4.2 we discuss the connections between our graphical estimation approach and methods from the wider deconfounding literature. If estimates of the covariance can be constructed by considering different spectral transformations (Ćevid et al., 2020), then since our procedure requires only a covariance estimate from the data, such a generalisation would be immediately compatible with out approach. However, further work is required to explore whether, and if so under what conditions, covariance estimation under a wider range of spectral transformations is valid.

4.7 Generating random graphs with v-structures

In the experiments of Section 4.5 we include examples where we guarantee the inclusion of some v-structures (see Figure 4.1). In practice, this means that there are some edges (j, k) with $\Omega_{jk} \neq 0$ but with $\Sigma_{jk} \approx 0$.

While it is straightforward to see how v-structures can (and do) arise within a SEM, for simplicity we have instead based our experimental settings on familiar examples of inverse covariance structures: banded, stochastic block model, Erdős-Rényi. Therefore in order to add-in v-structures, we construct a structural equation model that is consistent with a given precision matrix Ω.

We begin by considering a linear Gaussian SEM, supposing without loss of generality that our topological ordering is simply the ordering of the variables, so $j \to k \implies j < k$. Then define $A \in \mathbb{R}^{p\times p}$ to be a matrix such that $X_j = \sum_{k<j} A_{kj} X_k + \varepsilon_j$, where $\varepsilon \sim \mathcal{N}_p(0, D)$ with $D \in \mathbb{R}^{p\times p}$ a diagonal matrix. Note that A is a strictly upper triangular matrix, since the ordering of the variables is a topological ordering over the nodes in the DAG. This gives us the following formula:

$$\begin{aligned} X &= A^T X + \varepsilon \\ \implies (I_p - A^T)X &= \varepsilon \\ \implies X &= (I_p - A^T)^{-1}\varepsilon \end{aligned}$$

noting that $(I_p - A^T)^{-1}$ exists and is lower triangular. This in turn implies that

$$\begin{aligned} \Sigma = \operatorname{cov}(X) = \operatorname{cov}\left((I_p - A^T)^{-1}\varepsilon\right) = (I_p - A^T)^{-1} D (I_p - A)^{-1}, \\ \implies \Omega = \Sigma^{-1} = (I_p - A) D^{-1} (I_p - A)^T. \end{aligned} \tag{4.19}$$

Thus, returning to the setting when we are given Ω, if we can find a permutation matrix Q such that $Q\Omega Q^T = UD^{-1}U^T$, where:

- U is upper triangular with 1s on the diagonal
- D is diagonal and with strictly positive entries

then we can interpret $A = I_p - U$ as the adjacency matrix and D the covariance matrix that together specify the model. Fortunately, this can be done straightforwardly using standard software for performing LU decompositions (which with a trivial modification can instead provide a 'UL' decomposition). This has the functionality of simultaneously searching for a 'fill-reducing' permutation Q, ensuring that the structural equation model is sparse and thus specifying a topological ordering.

With this adjacency matrix, we simply scan through unconnected pairs (j, k) and independently with probability ρ_v impute $A_{jl} = A_{kl} = c$, for some specified constant c and randomly chosen $l > \max\{j, k\}$ such that $j \not\to l$ and $k \not\to l$. Columns of A are then rescaled as necessary so that any non-zero column sums are held constant. With this, we then use the formula (4.19) to reconstruct the precision matrix, which will still be sparse.

4.8 Proofs

4.8.1 Proof of Theorem 4.3

Proof. This proof follows the same basic structure as that of Theorem 5 in Shah et al. (2020), modified for the screening setting. We first note that from Theorem 3 of the same paper, there exists $\kappa > 0$ such that $\|\Sigma - \kappa\hat{\Sigma}^{\text{rsvp}}\|_\infty \lesssim \sqrt{\log(p)/n}$ with probability at least $1 - c/p$ for some constant c. In particular, combining Lemma 8 and equation (37) in its supplement gives that $\kappa \asymp p/n$.

We will work on the event that for each $j = 1, \ldots, p$, $\|\beta^j_{\hat{S}_j^c}\|_1 \lesssim \sqrt{\log(p)/n}$. We begin by rewriting our model for the j^{th} variable:

$$\begin{aligned} \boldsymbol{X}_j &= \boldsymbol{X}_{-j}\beta^j + \varepsilon^j \\ &= \boldsymbol{X}_{\hat{S}}\gamma^j + \boldsymbol{X}_{-j}\delta^j + \varepsilon^j, \end{aligned} \tag{4.20}$$

where note we lose the j subscript for $\hat{S}_j$. The population minimiser $\beta^j = \Sigma_{-j,-j}^{-1}\Sigma_{-j,j}$ is decomposed $\beta^j = \gamma^j + \delta^j$ such that $\gamma^j = \beta^j_{\hat{S}} \in \mathbb{R}^{|\hat{S}|}$, and $\delta^j \in \mathbb{R}^{p-1}$ is defined:

$$\delta^j_k = \begin{cases} \beta^j_k & \text{if } k \notin \hat{S}, \\ 0 & \text{if } k \in \hat{S}. \end{cases}$$

Note that by this definition, we have that $\|\delta^j\|_1 \lesssim \sqrt{\log(p)/n}$. Recall that $\varepsilon^j \sim \mathcal{N}_n(0, \Omega_{jj}^{-1} I_n)$, where $\Omega_{jj} \geq 1$ and note that $\varepsilon^j \perp\!\!\!\perp \boldsymbol{X}_{-j}$. Using now $\hat{\Sigma} = \hat{\Sigma}^{\text{rsvp}}$, our objective is

$$\hat{\beta}^j \in \operatorname*{arg\,min}_{b \in \mathbb{R}^{|\hat{S}|}} \frac{1}{2} b^T \hat{\Sigma}_{\hat{S},\hat{S}} b - \hat{\Sigma}_{j,\hat{S}} b + \lambda_j \|b\|_1, \tag{4.21}$$

for which the KKT conditions for $\hat{\beta}^j$ are:

$$\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}} \hat{\beta}^j = \lambda_j \hat{\nu}^j, \tag{4.22}$$

where $\|\hat{\nu}^j\|_\infty \leq 1$, and for each $k \in \hat{S}$, $\hat{\beta}^j_k \neq 0 \implies \hat{\nu}^j_k = \text{sgn}\hat{\beta}^j_k$. It is worth noting that in the above, we have $\left(\hat{\beta}^j\right)^T \hat{\nu}^j = \|\hat{\beta}^j\|_1$; for any general vector $v \in \mathbb{R}^{|\hat{S}|}$, we also have $v^T \hat{\nu}^j \leq \|v\|_1$. By taking an inner product with $\gamma^j - \hat{\beta}^j$ and rearranging, we arrive at the basic inequality

$$\begin{aligned} \left(\gamma^j - \hat{\beta}^j\right)^T \hat{\Sigma}_{\hat{S},\hat{S}} \left(\gamma^j - \hat{\beta}^j\right) &\leq \lambda_j \left(\|\gamma^j\|_1 - \|\hat{\beta}^j\|_1\right) + \left(\hat{\beta}^j - \gamma^j\right)^T \left(\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}} \gamma^j\right) \\ &\leq \lambda_j \left(\|\gamma^j\|_1 - \|\hat{\beta}^j\|_1\right) + \|\gamma^j - \hat{\beta}^j\|_1 \|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}} \gamma^j\|_\infty. \end{aligned} \tag{4.23}$$

We will now define the events

- $\Lambda^{(1)}_j = \{\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}} \gamma^j\|_\infty \leq \lambda_j/2\}$,
- $\Lambda^{(2)} = \{\|\kappa\hat{\Sigma} - \Sigma\|_\infty \leq (1 \wedge \sigma_l/32s)\}$, very similar to the quantity $(\Lambda^{(2)})$ defined in the proof of Theorem 5 in Shah et al. (2020) (there it appears without the $1 \wedge$),

and will work on their intersection, returning at the end of the proof to show this intersection holds with high probability. Since we are working on $\Lambda^{(1)}_j$, we can obtain

$$\left(\gamma^j - \hat{\beta}^j\right)^T \hat{\Sigma}_{\hat{S},\hat{S}} \left(\gamma^j - \hat{\beta}^j\right) \leq \lambda_j \left(\|\gamma^j\|_1 - \|\hat{\beta}^j\|_1\right) + \frac{\lambda_j}{2} \|\gamma^j - \hat{\beta}^j\|_1. \tag{4.24}$$

The rest of the proof follows the standard approach for bounding estimation error for the Lasso (see e.g. Chapter 6 of Bühlmann and van de Geer (2011)). Denoting $S = S_j \cap \hat{S}$, the set of true positives, we can observe that

$$2\|\gamma^j_S\|_1 - 2\|\hat{\beta}^j_S\|_1 - 2\|\hat{\beta}^j_{S^c}\|_1 + \|\gamma^j_S - \hat{\beta}^j_S\|_1 + \|\hat{\beta}^j_{S^c}\|_1 \geq 0.$$

Rearranging and applying the triangle inequality, we obtain

$$3\|\gamma^j_S - \hat{\beta}^j_S\|_1 \geq \|\gamma^j_{S^c} - \hat{\beta}^j_{S^c}\|_1. \tag{4.25}$$

The compatibility constant $\phi^2(M, S')$ is defined for some set S', some matrix M,:

$$\phi^2(M, S') = \min_{b:3\|b_{S'}\|_1 \geq \|b_{S'^c}\|_1} \frac{b^T M b}{\|b\|_2^2}. \tag{4.26}$$

If $\phi^2_j(M, S') > 0$, we say the *compatibility condition* is satisfied. If we have two sets $S_1 \subseteq S_2$, we can see that $\phi^2(M, S_1) \geq \phi^2(M, S_2)$. Similarly, if M' is a principal submatrix of M then $\phi^2(M', S') \geq \phi^2(M, S')$.

The application of Corollary 6.8 in Bühlmann and van de Geer (2011) gives that if $\|\hat{\Sigma}_{-j,-j} - \Sigma_{-j,-j}\|_\infty \leq \phi^2(\Sigma_{-j,-j}, S)/32|S_j|$ then $\phi^2(\hat{\Sigma}_{-j,-j}, S) \geq \phi^2(\Sigma_{-j,-j}, S)/2$. This will be the case on the event $\Lambda^{(2)}$. It then follows from the definition (4.26) that $\phi^2(\Sigma_{-j,-j}, S_j) \geq \sigma_l$. We define $\phi^2_j = \phi^2(\Sigma_{-j,-j}, S_j)$, noting that this definition involves the full support set $S_j \supseteq S$. We then have

$$\sigma_l \leq \phi^2_j \leq 2\phi^2(\hat{\Sigma}_{\hat{S},\hat{S}}, S), \tag{4.27}$$

which we use to obtain the final result.

By rearranging (4.24), and using (4.25), we obtain

$$\begin{aligned}
\left(\gamma^j - \hat{\beta}^j\right)^T \hat{\Sigma}_{\hat{S},\hat{S}} \left(\gamma^j - \hat{\beta}^j\right) + \frac{\lambda_j}{2}\|\gamma^j - \hat{\beta}^j\|_1 &\leq 8\lambda_j \|\gamma^j_S - \hat{\beta}^j_S\|_1 \\
&\leq 8\lambda_j \sqrt{|S|}\|\gamma^j_S - \hat{\beta}^j_S\|_2 \\
&\leq 8\lambda_j \sqrt{|S|}\|\gamma^j - \hat{\beta}^j\|_2 \\
&\leq 8\lambda_j \frac{\sqrt{|S|}\sqrt{(\gamma^j - \hat{\beta}^j)^T \hat{\Sigma}_{\hat{S},\hat{S}}(\gamma^j - \hat{\beta}^j)}}{\phi(\hat{\Sigma}_{\hat{S},\hat{S}}, S)} \\
&\leq 16\lambda_j \frac{\sqrt{|S_j|}\sqrt{(\gamma^j - \hat{\beta}^j)^T \hat{\Sigma}_{\hat{S},\hat{S}}(\gamma^j - \hat{\beta}^j)}}{\phi_j},
\end{aligned} \tag{4.28}$$

from which we are able to obtain

$$\left(\gamma^j - \hat{\beta}^j\right)^T \hat{\Sigma}_{\hat{S},\hat{S}} \left(\gamma^j - \hat{\beta}^j\right) + \frac{\lambda_j}{2}\|\gamma^j - \hat{\beta}^j\|_1 \leq \frac{64\lambda_j^2 |S_j|}{\phi_j^2}.$$

This gives the results

$$\|\gamma^j - \hat{\beta}^j\|_1 \leq \frac{128|S_j|A\sqrt{\log(p)}}{\sqrt{n}\sigma_l} \lesssim s_j\sqrt{\frac{\log(p)}{n}} \tag{4.29}$$

$$\|\beta^j - \hat{\beta}^j\|_1 \lesssim s_j\sqrt{\frac{\log(p)}{n}}, \tag{4.30}$$

where the second line comes by the triangle inequality and the assumption that $\|\delta^j\|_1 \leq \lambda_j/8$. Note we abuse notation in (4.30) by referring to a vector $\hat{\beta}^j \in \mathbb{R}^{p-1}$, which matches the $\hat{S}$-dimensional object on $\hat{S}$ and has zeroes at indices $k \notin \hat{S}$.

By the definition (4.26) of the compatibility constant, we also have that $\|\gamma^j - \hat{\beta}^j\|_2^2 \leq \frac{2}{\phi_j^2}(\gamma^j - \hat{\beta}^j)^T\hat{\Sigma}_{\hat{S},\hat{S}}(\gamma^j - \hat{\beta}^j)$, which combines with the results above to give:

$$\|\gamma^j - \hat{\beta}^j\|_2 \leq \frac{8\sqrt{2}A\sqrt{|S_j|\log(p)}}{\sqrt{n}\sigma_l} \lesssim \sqrt{\frac{s_j\log(p)}{n}}$$

$$\|\beta^j - \hat{\beta}^j\|_2 \lesssim \sqrt{\frac{s_j\log(p)}{n}}.$$

We now consider the events $\Lambda_j^{(1)}$ and $\Lambda^{(2)}$ and show that their intersection holds with probability at least $1 - c/p$, for some constant $c > 0$. To control $\Lambda_j^{(1)}$, we observe that

$$\begin{aligned}\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty &\leq \|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},-j}\beta^j\|_\infty + \|\hat{\Sigma}_{\hat{S},-j}\delta^j\|_\infty \\ &\leq \|\hat{\Sigma}_{-j,j} - \hat{\Sigma}_{-j,-j}\beta^j\|_\infty + \|\hat{\Sigma}_{-j,-j}\delta^j\|_\infty \\ &\leq \|\hat{\Sigma}_{-j,j} - \hat{\Sigma}_{-j,-j}\beta^j\|_\infty + \|(\hat{\Sigma}_{-j,-j} - \kappa^{-1}\Sigma_{-j,-j})\delta^j\|_\infty + \kappa^{-1}\|\Sigma_{-j,-j}\delta^j\|_\infty.\end{aligned} \tag{4.31}$$

Being careful not to get mixed up with the notation, the first term in (4.31) is exactly to the quantity controlled under the event denoted ($\Lambda_j^{(1)}$) in the proof of Theorem 5 in Shah et al. (2020). Appealing to this result gives that $\|\hat{\Sigma}_{-j,j} - \hat{\Sigma}_{-j,-j}\beta^j\|_\infty \leq \lambda_j/4$ for each j provided A is sufficiently large, with probability at least $1 - c/p$ for some constant $c > 0$.

For controlling the second and third terms in (4.31), we use Hölder's inequality:

$$\|(\hat{\Sigma}_{-j,-j} - \kappa^{-1}\Sigma_{-j,-j})\delta^j\|_\infty + \kappa^{-1}\|\Sigma_{-j,-j}\delta^j\|_\infty \leq \left(\|\hat{\Sigma}_{-j,-j} - \kappa^{-1}\Sigma_{-j,-j}\|_\infty + \kappa^{-1}\|\Sigma_{-j,-j}\|_\infty\right)\|\delta^j\|_1.$$

We already know that $\|\Sigma_{-j,-j}\|_\infty = 1$ and by Theorem 3 in Shah et al. (2020), we have that with probability at least $1 - c/p$ for some constant c that $\|\kappa\hat{\Sigma} - \Sigma\|_\infty \lesssim \sqrt{\log(p)/n}$, and since $s = o(\sqrt{\log(p)/n})$ and $\sqrt{\log(p)/n} = o(1)$ we have that $\Lambda^{(2)}$ holds with probability at least

$1 - c/p$. Therefore for sufficiently large n, on $\Lambda^{(2)}$,

$$\|\kappa\hat{\Sigma} - \Sigma\|_\infty + \|\Sigma\|_\infty \leq 2.$$

Since we are working on the event that $\|\delta^j\|_1 \lesssim \sqrt{\log(p)/n}$, it follows from $\kappa \gtrsim p/n$ that for sufficiently large n and A that:

$$\left(\|\hat{\Sigma}_{-j,-j} - \kappa^{-1}\Sigma_{-j,-j}\|_\infty + \kappa^{-1}\|\Sigma_{-j,-j}\|_\infty\right)\|\delta^j\|_1 \leq \frac{\lambda_j}{4}.$$

It follows that for sufficiently large n and A, with probability at least $1 - c/p$ for some constant c, that for $j = 1, \ldots, p$,

$$\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty \leq \frac{\lambda_j}{2}.$$

□

4.8.2 Proof of Theorem 4.3 with empirical covariance estimate

Proof. This proof closely follows the lines of the proof of Theorem 4.3 for the setting with confounding. We now refer to $\hat{\Sigma} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$. Our objective is:

$$\hat{\beta}^j \in \underset{\beta\in\mathbb{R}^{|\hat{S}|}}{\arg\min}\left\{\frac{1}{2}\beta^T\hat{\Sigma}_{\hat{S},\hat{S}}\beta - \hat{\Sigma}_{j,\hat{S}}\beta + \lambda_j\|\beta\|_1\right\}. \tag{4.32}$$

Identically to before, we decompose $\beta^j = \gamma^j + \delta^j$ and can obtain the same inequality as (4.23):

$$\left(\gamma^j - \hat{\beta}^j\right)^T\hat{\Sigma}_{\hat{S},\hat{S}}\left(\gamma^j - \hat{\beta}^j\right) \leq \lambda_j\left(\|\gamma^j\|_1 - \|\hat{\beta}^j\|_1\right) + \|\gamma^j - \hat{\beta}^j\|_1\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty.$$

We now deviate very slightly from the previous proof of Theorem 4.3 by defining the following events, the intersection of which we will work on:

- $\Lambda_j^{(1)} = \{\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty \leq \lambda_j/2\}$
- $\Lambda^{(2)} = \{\|\hat{\Sigma} - \Sigma\|_\infty \leq (1 \wedge \sigma_l/32s)\}$

On $\Lambda_j^{(1)}$, we have as before that:

$$\left(\gamma^j - \hat{\beta}^j\right)^T\hat{\Sigma}_{\hat{S},\hat{S}}\left(\gamma^j - \hat{\beta}^j\right) \leq \lambda_j\left(\|\gamma^j\|_1 - \|\hat{\beta}^j\|_1\right) + \|\gamma^j - \hat{\beta}^j\|_1\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty.$$

By the same argument as for obtaining (4.27), and defining $\phi_j^2 = \phi^2(\Sigma_{-j,-j}, S_j)$, it follows that on $\Lambda^{(2)}$ we have

$$\sigma_l \leq \phi_j^2 \leq 2\phi^2(\hat{\Sigma}_{\hat{S},\hat{S}}, S).$$

We then follow the same steps until (4.28), where we obtain as before:

$$(\gamma^j - \hat{\beta}^j)^T \hat{\Sigma}_{\hat{S},\hat{S}}(\gamma^j - \hat{\beta}^j) + \frac{\lambda_j}{2}\|\gamma^j - \hat{\beta}^j\|_1 \leq \frac{64\lambda_j^2|S_j|}{\phi_j^2}.$$

This first gives the results

$$\|\gamma^j - \hat{\beta}^j\|_1 \leq \frac{128|S_j|A\sqrt{\log(p)}}{\sqrt{n}\sigma_l} \lesssim s_j\sqrt{\frac{\log(p)}{n}} \tag{4.33}$$

$$\|\beta^j - \hat{\beta}^j\|_1 \lesssim s_j\sqrt{\frac{\log(p)}{n}}, \tag{4.34}$$

where the second line comes by the triangle inequality and the assumption that $\|\delta^j\|_1 \leq \lambda_j/8$. Note we abuse notation in (4.34) by referring to a vector $\hat{\beta}^j \in \mathbb{R}^{p-1}$ which simply has zero entries on indices $k \notin \hat{S}$.

By the definition of the compatibility constant, we also have that $\|\gamma^j - \hat{\beta}^j\|_2^2 \leq \frac{2}{\phi_j^2}(\gamma^j - \hat{\beta}^j)^T\hat{\Sigma}_{\hat{S},\hat{S}}(\gamma^j - \hat{\beta}^j)$, which combines with the results above to give:

$$\|\gamma^j - \hat{\beta}^j\|_2 \leq \frac{8\sqrt{2}A\sqrt{|S_j|\log(p)}}{\sqrt{n}\sigma_l} \lesssim \sqrt{\frac{s_j\log(p)}{n}}$$

$$\|\beta^j - \hat{\beta}^j\|_2 \lesssim \sqrt{\frac{s_j\log(p)}{n}}.$$

We will now turn out attention to the events $\Lambda_j^{(1)}$ and $\Lambda^{(2)}$. We will first show that with high probability, $\Lambda_j^{(1)}$ holds for each j.

Using (4.20), it by using the triangle inequality that

$$\begin{aligned}\|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},\hat{S}}\gamma^j\|_\infty &\leq \|\hat{\Sigma}_{\hat{S},j} - \hat{\Sigma}_{\hat{S},-j}\beta^j\|_\infty + \|(\hat{\Sigma}_{\hat{S},-j} - \Sigma_{\hat{S},-j})\delta^j\|_\infty + \|\Sigma_{\hat{S},-j}\delta^j\|_\infty \\ &\leq \overbrace{\|\hat{\Sigma}_{-j,j} - \hat{\Sigma}_{-j,-j}\beta^j\|_\infty}^{\text{(I)}} + \overbrace{\|(\hat{\Sigma}_{-j,-j} - \Sigma_{-j,-j})\delta^j\|_\infty}^{\text{(II)}} + \overbrace{\|\Sigma_{-j,-j}\delta^j\|_\infty}^{\text{(III)}}.\end{aligned}$$

To control (I), we have that $\|\hat{\Sigma}_{-j,j} - \hat{\Sigma}_{-j,-j}\beta^j\|_\infty = \frac{1}{n}\|\boldsymbol{X}_{-j}^T\varepsilon^j\|_\infty$. First conditioning on $\boldsymbol{X}_{-j}$, the l^{th} component of $\boldsymbol{X}_{-j}^T\varepsilon^j$ is a Gaussian random variable[3] with variance $\Omega_{jj}^{-1}\|\boldsymbol{X}_l\|_2^2$, inviting the tail bound

$$\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{X}_{-j}^T\varepsilon^j\|_\infty > \frac{\lambda_j}{4}\middle|\boldsymbol{X}_{-j}\right) < 2(p-1)\max_{l\neq j}\exp\left(-\frac{n^2\lambda_j^2\Omega_{jj}}{32\|\boldsymbol{X}_l\|_2^2}\right).$$

Then since $\|\boldsymbol{X}_l\|_2^2 \preceq Z$ for some random variable $Z \sim \chi_n^2$ (since $\boldsymbol{X}_l$ has been centred), we have $\mathbb{P}(\max_{l\neq j}\|\boldsymbol{X}_l\|_2^2/n \geq 2) \leq (p-1)\exp\left(-\frac{4-\sqrt{12}}{4}n\right)$ using the chi-squared bound of Lemma 1 in Laurent and Massart (2000). Therefore we have that

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\|\boldsymbol{X}_{-j}^T\varepsilon^j\|_\infty \leq \frac{\lambda_j}{4}\right) &\geq 1-(p-1)\left(2\exp\left(-\frac{n\lambda_j^2}{64}\right)+\exp\left(-\frac{4-\sqrt{12}}{4}n\right)\right)\\ &\geq 1-2p^{-\left(\frac{A^2}{64}-1\right)}-p^{-(B-1)},\end{aligned}$$

where we use that $\log(p)/n \to 0$ so for sufficiently large n, $n > 4B\log(p)/(4-\sqrt{12})$. It follows that

$$\mathbb{P}\left(\bigcap_j\left\{\frac{1}{n}\|\boldsymbol{X}_{-j}^T\varepsilon^j\|_\infty \leq \frac{\lambda_j}{4}\right\}\right) \geq 1-2p^{-\left(\frac{A^2}{64}-2\right)}-p^{-(B-1)}$$

Let us now consider the event $\Lambda^{(2)}$. The standard approach for bounding $\|\Sigma-\hat{\Sigma}\|_\infty$ uses Bernstein's inequality on the product of sub-Gaussian tail bounds, which in our specific (Gaussian) case gives the result

$$\mathbb{P}(\|\hat{\Sigma}-\Sigma\|_\infty > t) < 2\exp(-nt^2/(128+8t)+2\log(p)). \tag{4.35}$$

Setting $t = \sigma_l/32s$, we have for a constant $C > 0$ that for sufficiently large[4] n,

$$\mathbb{P}\left(\|\hat{\Sigma}-\Sigma\|_\infty \leq \frac{\sigma_l}{32s}\right) \geq \mathbb{P}\left(\|\hat{\Sigma}-\Sigma\|_\infty \leq \sqrt{257+C}\sqrt{\frac{\log(p)}{n}}\right) \geq 1-2p^{-C}.$$

Together, the terms (II) and (III) can be bounded

$$(\text{II})+(\text{III}) \leq \left(\|\hat{\Sigma}_{-j,-j}-\Sigma_{-j,-j}\|_\infty + \|\Sigma_{-j,-j}\|_\infty\right)\|\delta^j\|_1.$$

[3]Recalling our assumption that all columns of $\boldsymbol{X}$ have been projected onto the subspace orthogonal to $\mathbf{1}$, it is sufficient to assume for the purposes of bounding $\|\boldsymbol{X}_{-j}^T\varepsilon^j\|_\infty$ (by the idempotence of projection matrices) that the columns of $\boldsymbol{X}_{-j}$ have been centred but ε^j has not.

[4]This is true for any $C > 0$, from our assumption that $s^2\log(p)/n \to 0$.

We already know that $\|\Sigma_{-j,-j}\|_\infty = 1$. Now working on $\Lambda^{(2)}$, we also have that $\|\hat{\Sigma} - \Sigma\|_\infty \leq 1$, so we have (II) + (III) $\leq 2\|\delta^j\|_1$. By assumption, $\|\delta^j\|_1 \leq \lambda_j/8$, so it now follows that the events $\cap_j \Lambda_j^{(1)}$ and $\Lambda^{(2)}$ hold with probability at least $1 - 2p^{-\left(\frac{A^2}{64}-2\right)} - p^{-(B-1)} - 2p^{-C}$.

□

4.8.3 Other proofs

Proof of Lemma 4.1. We note that while the scaling factor κ of the estimated covariance may unknown, Algorithm 4.2 relies only on estimated correlation and partial correlation statistics from $\hat{\Sigma}$, both of which depend only on the correlation matrix obtained by rescaling the rows and columns of $\hat{\Sigma}$ (Kim, 2015). Thus they are both invariant under scaling of the covariance estimate used.

To prove part (a), we establish that the correlation and partial correlation tests performed during Algorithm 4.2 are sufficiently sensitive. Provided $\|\kappa\hat{\Sigma} - \Sigma\|_\infty < \Xi$ for some $\kappa > 0$, it follows that

$$\hat{\rho}_{jk} \in [(1-\Xi)(\rho_{jk} - \Xi), (1-\Xi)^{-1}(\rho_{jk} + \Xi)]. \tag{4.36}$$

Thus, if $|\rho_{jk}| \geq \omega_1$, it follows that $\hat{\rho}_{jk} \geq (1-\Xi)(\omega_1 - \Xi)$. Therefore provided our threshold $\tau_1 \leq (1-\Xi)(\omega_1 - \Xi)$, we have that $(|\rho_{jk}| \geq \omega_1) \implies (|\hat{\rho}_{jk}| \geq \tau_1)$ and thus $k \in S_j^{\text{init}}$ (and also $j \in S_k^{\text{init}}$).

It follows from the application of Lemma 4 in Harris and Drton (2013) to the matrix $\left(\kappa\hat{\Sigma}_{\{j,k,l\},\{j,k,l\}} - \Sigma_{\{j,k,l\},\{j,k,l\}}\right)$, that $|\rho_{jk|l} - \hat{\rho}_{jk|l}| \leq \omega_2/4$. Therefore, provided our threshold $\tau_2 \leq 3\omega_2/4$, we have that $(\rho_{jk|l}| \geq \omega_2) \implies (|\hat{\rho}_{jk|l}| \geq 3\omega_2/4)$ and therefore provided $\{j,l\} \subseteq S_k^{\text{init}}$ then $k \in \hat{S}_j$ and $j \in \hat{S}_k$. The first claim now follows straightforwardly.

We now turn our attention to part (b). We have that for $k \notin S_j \cup \mathcal{F}_j^{(1)}(\tau_1, \Xi)$ that $|\rho_{jk}| < (1+\tau_1)(1-\Xi) - 1$. By (4.36), it follows that $|\hat{\rho}_{jk}| < \tau_1$ and therefore $k \notin S_j^{\text{init}}$ and $j \notin S_k^{\text{init}}$. Similarly, since for $k \notin S_j \cup \mathcal{F}_j^{(2)}(\tau_2, \omega_2)$ we have that $\max_{k \in V\setminus\{j,k\}} |\rho_{jk|l}| < \tau_2 - \omega_2/4$. In order for $k \in \hat{S}_j \setminus S_j$, we need either that $k \in S_j^{\text{init}}$, or that there exists $l \in V\setminus\{j,k\}$ such that $l \in S_j^{\text{init}}$ and $l \in S_k^{\text{init}}$, and $|\hat{\rho}_{jk|l}| \geq \tau_2$. The first case can only happen when $k \in \mathcal{F}_j^{(1)}(\tau_1, \Xi)$ and the second can only happen when $k \in \mathcal{F}_j^{(2)}(\tau_2, \omega_2)$. This concludes the proof.

□

Proof of Theorem 4.2. This follows as a consequence of Theorem 4.3 and Lemma 4.1, noting that on the event $\Lambda^{(2)}$ in the relevant version of the proof of Theorem 4.3 we have that (4.6) is satisfied. Lemma 1 then implies that $\|\beta^j_{\hat{S}_j^c}\|_1 = 0$ for $j = 1, \ldots, p$. □

4.9 Supplementary discussion of Iterative SIS

We have discussed in Section 4.2.1 why our assumptions are reasonable and allow the construction of a two-step screening procedure that possesses the sure screening property in our graphical estimation setting. For general regression problems this is not possible without making further assumptions, as there exist examples where important variables do not appear (in a marginal correlation sense) for an arbitrary number of iterations of a procedure such as we outline below.

Consider the iterative SIS (Fan and Lv, 2008) procedure, which for the linear model case can be described in the following steps:

1. Correlation set C are those variables in X sufficiently correlated with Y.
2. Regress Y on X_C
3. Include in correlation set C all those variables in X correlated with the residuals form the above regression. Repeat until stop.

Examples exist for which these steps need repeating an arbitrarily large number of times before all of the true variables have non-zero (population) correlation with the residuals. We construct one as follows (excluding any non-signal variables for simplicity):

- $M_1 = \begin{pmatrix} 1 \end{pmatrix}, \quad \beta^{(1)} = [1]$
- $M_{k+1} = \left(\begin{array}{cc|c} M_k & \mathbf{0} & \begin{matrix} 0 \\ \vdots \\ a_k \end{matrix} \\ \hline \mathbf{0} & M_k & \begin{matrix} 0 \\ \vdots \\ a_k \end{matrix} \\ \hline 0 \cdots a_k & 0 \cdots a_k & 1 \end{array}\right), \quad \beta^{(k+1)} = [\beta^{(k)}, \beta^{(k)}, (-1)^k]$
- The sequence (a_k) is defined as follows. The final term $a_{K-1} = \frac{1}{2}$, and then previous terms are defined $a_{k-1} = \frac{1-a_k}{2}$.

Theorem 1.21 in Varga (2000) states that if a matrix is diagonally dominant (strictly in at least one row) and irreducible (in a symmetric matrix, the graph with edges of nonzero entries is a single connected component) then the eigenvalues are strictly positive, implying that the M_k are indeed covariance matrices. It is easy to check the above matrices are irreducible, and diagonal dominance follows from the construction of the sequence a_k.

In this construction, the coefficients that are some multiples of $2^k - 1$ are uncorrelated with all except those which are multiples of $2^{k-1} - 1$. Essentially this means that each time

we regress on the residuals, some new variables are correlated with the residuals and we must repeat this K times to select all signal variables.

Bibliography

V. Aggarwala and B. F. Voight. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4): 349–355, 2016.

F. Aguet, A. A. Brown, S. E. Castel, J. R. Davis, Y. He, B. Jo, P. Mohammadi, Y. Park, P. Parsana, A. V. Segrè, B. J. Strober, Z. Zappala, N. S. Abell, B. Balliu, O. Delaneau, L. Frésard, E. R. Gamazon, D. Garrido-Martín, A. D. H. Gewirtz, G. Gliner, M. J. Gloudemans, B. Han, A. Z. He, F. Hormozdiari, X. Li, B. Liu, E. Y. Kang, I. C. McDowell, H. Ongen, J. J. Palowitch, C. B. Peterson, G. Quon, S. Ripke, A. Saha, A. A. Shabalin, T. C. Shimko, J. H. Sul, N. A. Teran, E. K. Tsang, H. Zhang, Y.-H. Zhou, C. D. Bustamante, N. J. Cox, R. Guigó, M. Kellis, M. I. McCarthy, D. F. Conrad, E. Eskin, G. Li, A. B. Nobel, C. Sabatti, B. E. Stranger, X. Wen, F. A. Wright, K. G. Ardlie, E. T. Dermitzakis, T. Lappalainen, GTEx Consortium, A. Battle, C. D. Brown, B. E. Engelhardt, and S. B. Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675): 204–213, 2017.

H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

E. Anderson. *rosqp: Quadratic Programming Solver using the 'OSQP' Library*, 2018. R package version 0.1.0.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, 5:19–53, 2011.

M. Barigozzi and H. Cho. Consistent estimation of high-dimensional factor models when the factor number is over-estimated. *Electronic Journal of Statistics*, 14(2):2892–2921, 2020.

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*, 2021.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103, 2015.

H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in anova. *Biometrics*, 65(1):169–177, 2009.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232, 2011.

P. Breheny and J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187, 2015.

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4): 373–384, 1995.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. The Wadsworth and Brooks-Cole statistics-probability series. CRC press, 1984.

P. Bühlmann and D. Ćevid. Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88:S114–S134, 2020.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer Berlin Heidelberg, 2011. ISBN 9783642201929.

T. Cai, W. Liu, and X. Luo. A constrained l-1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

T. Calinski and L. Corsten. Clustering means in anova by simultaneous testing. *Biometrics*, pages 39–48, 1985.

E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

D. Ćevid, P. Bühlmann, and N. Meinshausen. Spectral deconfounding via perturbed sparse linear models. *Journal of Machine Learning Research*, 21:232, 2020.

J. Chen and Z. Chen. Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

V. Chernozhukov, C. Hansen, and Y. Liao. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics (to appear)*, 2021.

J. Chiquet, P. Gutierrez, and G. Rigaill. Fast tree inference with weighted fusion penalties. *Journal of Computational and Graphical Statistics*, 26(1):205–216, 2017.

H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

A. Datta and H. Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values and r-software hdi. *Statistical Science*, pages 533–558, 2015.

D. Dua and C. Graff. UCI machine learning repository, 2019.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494): 544–557, 2011.

J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 2013.

J. Fan, H. Liu, Q. Sun, and T. Zhang. I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814, 2018.

Y. Feng and Y. Yu. The restricted consistency property of leave-nv-out cross-validation for high-dimensional variable selection. *Statistica Sinica*, 29(3):1607–1630, 2019.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

L. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

J. Gertheiss and G. Tutz. Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4):2150–2180, 2010.

I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

A. Gorst-Rasmussen and T. Scheike. Independent screening for single-index hazard rate models with ultrahigh dimensional features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):217–245, 2013.

N. Harris and M. Drton. Pc algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14(69):3365–3383, 2013.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009. ISBN 9780387848846.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2016.

T. D. Hocking, A. Joulin, F. Bach, and J.-P. Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th International Conference on Machine Learning*, page 1, 2011.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

S. Hu, A. O'Hagan, and T. B. Murphy. Motor insurance claim modelling with factor collapsing and bayesian model averaging. *Stat*, 7(1):e180, 2018.

J. Huang, Y. Jiao, Y. Liu, and X. Lu. A constructive approach to l_0 penalized regression. *Journal of Machine Learning Research*, 19:1–37, 2018.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):e12776, 2010.

L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 433–440, 2009.

W. James and C. M. Stein. pages 361–379. University California Press, Berkeley, 1961.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395, 2012.

J. Jia and K. Rohe. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172, 2015.

N. A. Johnson. A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Kaggle. Prudential life insurance assessment. https://www.kaggle.com/c/prudential-life-insurance-assessment. Accessed: 2021-08-25.

W. Karush. Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago, 1939.

J. Kiggins. Avocado prices. https://www.kaggle.com/neuromusic/avocado-prices. Accessed: 2021-06-04.

S. Kim. ppcor: an r package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods*, 22(6):665, 2015.

Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13(1):1037–1057, 2012.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. pages 481–492. University of California Press, 1951.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

S. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996. ISBN 9780191591228.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413, 2014.

P.-L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12(1):1429–1467, 2018.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

P.-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(1): 559–616, 2015.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

Y. Lu and H. H. Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

S. Luo, R. Song, and D. Witten. Sure screening for gaussian graphical models. *arXiv preprint arXiv:1407.7819*, 2014.

S. Ma and J. Huang. A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, 112(517):410–423, 2017.

A. Maj-Kańska, P. Pokarowski, and A. Prochenka. Delete or merge regressors for linear model selection. *Electronic Journal of Statistics*, 9(2):1749–1778, 2015.

R. Mazumder, J. H. Friedman, and T. Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

C. Meek. Strong completeness and faithfulness in bayesian networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 411—-418, 1995.

N. Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

J. Menchero, A. Morozov, and P. Shepard. Global equity risk modeling. In *Handbook of Portfolio Construction*, pages 439–480. Springer, 2010.

C. Micchelli, J. Morales, and M. Pontil. A family of penalty functions for structured sparsity. *Advances in Neural Information Processing Systems*, 23:1612–1623, 2010.

A. J. Miller. *Subset Selection in Regression*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1990. ISBN 9780412353802.

Y. Nardi and A. Rinaldo. Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, 102(3):528–549, 2011.

J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap, K. S. King, S. Bergmann, M. R. Nelson, et al. Genes mirror geography within europe. *Nature*, 456 (7218):98–101, 2008.

M.-R. Oelker, W. Pößnecker, and G. Tutz. Selection and fusion of categorical predictors with l 0-type penalties. *Statistical Modelling*, 15(5):389–410, 2015.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

D. Pauger and H. Wagner. Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, 14(2):341–369, 2019.

D. Pauger, M. Leitner, H. Wagner, and G. Malsiner-Walli. effectfusion: Bayesian effect fusion for categorical predictors, 2019.

J. Pearl. Probabilistic reasoning in intelligent systems; networks of plausible inference. Technical report, 1988.

A. Prochenka-Sołtys and P. Pokarowski. *DMRnet: Delete or Merge Regressors Algorithms for Linear and Logistic Model Selection and High-Dimensional Data*, 2018. R package version 0.2.0.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

J. Ramsey, P. Spirtes, and J. Zhang. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 401–408, 2006.

M. Rosenbaum and A. B. Tsybakov. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics, 2013.

K. Sadeghi. Faithfulness of probability distributions and graphs. *Journal of Machine Learning Research*, 18(148):1–29, 2017.

R. Samworth. Small confidence sets for the mean of a spherically symmetric distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):343–361, 2005.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.

R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75 (1):55–80, 2013.

R. D. Shah, B. Frot, G.-A. Thanei, and N. Meinshausen. Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):361–389, 2020.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

B. Stokell, D. Grose, and R. Shah. Catreg: Solution paths for linear and logistic regression models with categorical predictors, with scope penalty. https://cran.r-project.org/package=CatReg, 2021a. R package version 2.0.3.

B. G. Stokell, R. D. Shah, and R. J. Tibshirani. Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3):579–611, 2021b.

T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

COVID-19 Forecast Hub. The covid-19 forecast hub. https://covid19forecasthub.org, 2020.

G.-A. Thanei. *Computationally efficient inference for large-scale data*. PhD thesis, ETH Zürich, 2018.

T. Therneau and B. Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. R package version 4.1-15.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

R. Tibshirani and X. Suo. An ordered lasso and sparse time-lagged regression. *Technometrics*, 58(4):415–423, 2016.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111 (514):600–620, 2016.

A. N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

J. W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

G. Tutz and J. Gertheiss. Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200, 2016.

S. van de Geer. *Estimation and testing under sparsity*. Springer, 2016.

S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

R. S. Varga. *Matrix Iterative Analysis*. Springer-Verlag Berlin Heidelberg, second edition, 2000. ISBN 9783540663218.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

C. Wang and B. Jiang. An efficient admm algorithm for high dimensional precision matrix estimation via penalized quadratic loss. *Computational Statistics & Data Analysis*, 142: 106812, 2020.

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

X. Wang and C. Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (3):589–611, 2016.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics*, 42(6):2164, 2014.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Series in Probability and Statistics. Wiley, 1990. ISBN 9780471917502.

A. Wille and P. Bühlmann. Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, 5(1), 2006.

D. M. Witten, J. H. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

S. Xu, B. Jia, and F. Liang. Learning moral graphs in construction of high-dimensional bayesian networks for mixed data. *Neural Computation*, 31(6):1183–1214, 2019.

J. Yin and H. Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 5(4):2630, 2011.

Y. Yu, J. Bradic, and R. J. Samworth. Confidence intervals for high-dimensional cox models. *Statistica Sinica*, 31:243–267, 2021.

M. Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11:2261–2286, 2010.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 217–242, 2014.

C.-H. Zhang and T. Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 11 2012.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

T. Zhao, H. Liu, K. Roeder, J. Lafferty, and L. Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13 (1):1059–1062, 2012.

T. Zhao, H. Liu, and T. Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180–218, 2018.

X. Zheng and W.-Y. Loh. Consistent variable selection in linear models. *Journal of the American Statistical Association*, 90(429):151–156, 1995.

Z. Zhu, T. Wang, and R. J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

H. Zou and H. H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 37(4):1733, 2009.